High Performance Computing 2014/15
Technology Compass
IBM® Special
Table of Contents and Introduction

IBM Technical Computing ..................................................... 4
Accelerate Insights and Results ............................................... 6
IBM Application Ready Solutions ........................................... 10

Enterprise-Ready Cluster & Workload Management .................. 16
IBM Platform HPC .................................................................. 18
What’s New in IBM Platform LSF 8 ....................................... 22
IBM Platform MPI 8.1 ............................................................... 32

Scalable, Energy Efficient HPC Systems .................................... 38
IBM NeXtScale System ............................................................. 40

General Parallel File System (GPFS) ........................................ 46
Technologies That Enable the Management of Big Data .......... 48
What’s New in GPFS Version 3.5 ............................................ 64
GPFS Storage Server – a home for Big Data ............................. 66
More than 30 years of experience in Scientific Computing

1980 marked the beginning of a decade where numerous startups were created, some of which later transformed into big players in the IT market. Technical innovations brought dramatic changes to the nascent computer market. In Tübingen, close to one of Germany’s prime and oldest universities, transtec was founded.

In the early days, transtec focused on reselling DEC computers and peripherals, delivering high-performance workstations to university institutes and research facilities. In 1987, SUN/Sparc and storage solutions broadened the portfolio, enhanced by IBM/RS6000 products in 1991. These were the typical workstations and server systems for high performance computing then, used by the majority of researchers worldwide.

In the late 90s, transtec was one of the first companies to offer highly customized HPC cluster solutions based on standard Intel architecture servers, some of which entered the TOP500 list of the world’s fastest computing systems.

Thus, given this background and history, it is fair to say that transtec looks back upon a more than 30 years’ experience in scientific computing; our track record shows nearly 500 HPC installations. With this experience, we know exactly what customers’ demands are and how to meet them. High performance and ease of management – this is what customers require today. HPC systems are for sure required to peak-perform, as their name indicates, but that is not enough: they must also be easy to handle. Unwieldy design and operational complexity must be avoided or at least hidden from administrators and particularly users of HPC computer systems.

transtec HPC solutions deliver ease of management, both in the Linux and Windows worlds, and even where the customer’s environment is of a highly heterogeneous nature. Even the dynamical provisioning of HPC resources as needed does not constitute any problem, thus further leading to maximal utilization of the cluster.

transtec HPC solutions use the latest and most innovative technology. Their superior performance goes hand in hand with energy efficiency, as you would expect from any leading edge IT solution. We regard these basic characteristics.

In 2010, transtec entered into a strategic partnership with IBM, surely one of the biggest players in the HPC world with a very strong brand. The flexibility and long-year experience of transtec, combined with the power and quality of IBM HPC systems constitute a perfect symbiosis and provide customers with the most optimal HPC solution imaginable. IBM NeXtScale systems are highly optimized for HPC workload in datacenter environments, regarding performance, flexibility, and energy, space and cooling efficiency. Platform HPC and LSF are both enterprise-ready HPC cluster and workload management solutions and are widespread in all kinds of industrial HPC environments.

Your decision for a transtec HPC solution means you opt for most intensive customer care and best service in HPC. Our experts will be glad to bring in their expertise and support to assist you at any stage, from HPC design to daily cluster operations, to HPC Cloud Services.

Last but not least, transtec HPC Cloud Services provide customers with the possibility to have their jobs run on dynamically provided nodes in a dedicated datacenter, professionally managed and individually customizable. Numerous standard applications like ANSYS, LS-Dyna, OpenFOAM, as well as lots of codes like Gromacs, NAMD, VMD, and others are pre-installed, integrated into an enterprise-ready cloud management environment, and ready to run.

Have fun reading the transtec HPC Compass 2014/15 IBM Special!
All companies want more compute power, faster networks, better access to data and applications available everywhere and at all times. But simply deploying a fast computer without considering all the elements involved in planning, deployment, installation, administration and ongoing maintenance and updates can actually hobble an organization, affecting productivity and possibly damaging profitability and brand value.

Today’s technical computing solutions require an end-to-end view of the hardware, operating environment, applications, data management, software and services. It is about the overall system – where purpose-built systems reside next to general purpose solutions to form a compute and systems capability that meets the most demanding requirements. Whether dealing with real-time trading systems, managing the smart grid, optimizing realtime customer relationship management across multiple distribution channels, or running computationally demanding electronic design automation workloads, you need more than one size fits all to tackle your unique challenges.
Analyzing big data requires more than fast processors
As technical computing moves towards a data centric model, the ability to deal with large sets of fast-moving structured and unstructured data becomes paramount. Whether analyzing market data to make critical business decisions, or running data-intensive simulations to better understand physical phenomena, analytic processes must be carried out in ever-shorter time spans to be of value to an organization. Generating insights from the exploding volume, velocity and variety of data requires optimized systems specifically architected for that task. To maximize performance, systems optimization must be done at every layer of the technology stack to exploit unique processor, memory and storage characteristics. The increasingly sophisticated technical computing workflows require computing tuned to domain knowledge and workload characteristics, hardware with multi core architectures and advanced threading, and software tuned from the operating system through the middleware stack.

Meeting the challenges of your particular operation
Multi-step, big data analytics also requires optimized workflows – which means organizations can no longer pick a technical computing solution based on a single benchmark, such as a server’s maximum processing power or its ability to run a particular workload faster than a competitor’s solution. Companies need to examine the various tasks in their big data analytics workflows and match the requirements with suitable technical computing solutions.
A broad portfolio of superior and innovative products and technology
The IBM vision for technical computing is to bring together technology, science, management and innovation to enable major improvements in business and society – and help build a smarter planet. IBM provides an extensive selection of technical computing options from a portfolio of servers, storage, software, services and financing components backed by access to subject matter experts and world-class support. IBM solutions can help you optimize workloads and overcome obstacles to parallelism and other revolutionary approaches to supercomputing.

The sky’s the limit
IBM recognizes that one size does not fit all. That’s why we introduced the IBM Engineering Solutions for Cloud. Based on proven IBM technology, Engineering Solutions for Cloud let organizations build a centralized, shared product development center that supports both interactive and batch design workloads. This solution enables designers and engineers to access the Technical Computing cloud environment from a laptop practically anywhere in the world, using interactive applications with 2D or 3D remote visualization significantly saving cost and minimizing the amount of data that must be transferred to and from the cloud.

Helping a broad array of industries
IBM is helping companies and organizations in more than a dozen industries. IBM has powerful, innovative solutions to companies’ most challenging and complex problems, that allow businesses and researchers to innovate, make critical technical and business decisions, achieve breakthrough results, and establish sustainable competitive advantage.

Making sense of dollars
Financial services firms are rethinking their strategies as they respond to the sweeping changes in the markets and the regulatory environment, and an incessant blizzard of data. In fact, some financial organizations consume market data at rates exceeding one million messages per second, twice the peak rates they experienced only a year ago. With an optimized technical computing system from IBM, financial services enterprise can process vast amounts of structured and unstructured data, in real time which means you won’t be lost in a data whiteout.

Engineering a smarter planet
Meeting the demands of today’s automotive, aerospace, defense and manufacturing engineers requires unprecedented computing power for structural analysis, noise, vibration, and harshness tests, crash analytics, and fluid dynamics. IBM offers computer aided engineering (CAE) optimized solutions that include systems, storage and software, from leading ISVs to help you streamline your development environment, reduce design-cycle times and infrastructure costs, and meet aggressive time-to-market deadlines.

Searching for black gold
A tectonic shift is underway in upstream petroleum computing. Reservoir modeling and sensor field data now interact in near real-time to dramatically improve the fidelity of the analysis, its accuracy, and reliability. With IBM technical computing systems, energy companies can reduce the duration and cost of problem solving in reservoir optimization and seismic imaging, continuing to advance the field of exploration and production.
Community collaboration
Complete technical computing solutions may require components supplied by specialized vendor such as applications and tools from ISVs, hardware for interconnection and acceleration of processing nodes, and state-of-the-art cooling technology for greener operation, to name a few. IBM maintains technical and business relationships with all the leading technical computing providers.

IBM also works closely with industry, open standards consortia, and government agencies around the world to facilitate technology advancement and deployment, and collaborates with leading academic institutions through our shared university research programs and fellowships. Such collaborations drive value back to the community and result in improved products.

Using insight to help support a smarter planet
Whether optimizing traffic flow to lowering fuel consumption and time wasted in traffic jams, or unraveling genetic codes to develop new medicines and therapies, or increasing the production of oil and gas from existing reservoirs, powerful and efficient technical computing solutions from IBM provide a foundation to handle the associated computational challenges and extract intelligence from complex systems of instrumented and interconnected people and devices.
Accelerate Time to Value for Technical Computing

Businesses in nearly every industry are looking for ways to improve the efficiency of their technical computing environments. Companies that design aerospace or automotive products need systems that can help them meet time-to-market requirements and maximize profitability. Organizations helping to find causes and cures for disease need ways to increase productivity, foster innovation and compete more effectively. For reservoir engineers, the rising cost of oil and gas drilling means evermore accurate models are required to pinpoint potential well sites and extract higher percentages of oil and gas resources, and communications service providers need a way to quickly analyze data and act on it.

Keys to overcoming technical computing challenges

Most technical computing tasks involve vast amounts of data and require thousands of complex calculations. Pressures to “do more with less” create requirements for greater efficiency. Increasing application performance and workload throughput is important—but that is only part of the solution. Organizations can also realize dramatic efficiency benefits from simplified installation, deployment and management of an optimized technical computing environment.

Additionally, many companies have limited IT resources to devote to administering the high-performance systems required for sophisticated design, analytics and research tasks. These companies require a solution that is affordable and easy to use, and that will help make the most of their infrastructure investment by ensuring compute resources are fully utilized and prioritized.
IBM has created workload-optimized solutions designed to meet these challenges. IBM Application Ready Solutions for Technical Computing are based on IBM Platform Computing software and powerful IBM systems, integrated and optimized for leading applications and backed by reference architectures. With IBM Application Ready Solutions, organizations can spend more time solving scientific and engineering problems, instead of administering computing environments.

**IBM Application Ready Solutions: Looking under the hood**

IBM has created Application Ready Solution reference architectures for target workloads and applications. Each of these reference architectures includes recommended small, medium and large configurations designed to ensure optimal performance at entry-level prices. These reference architectures are based on powerful, predefined and tested infrastructure with a choice of the following systems:

- IBM Flex System provides the ability to combine leading-edge compute nodes with integrated storage and networking in a highly dense, scalable blade system. The IBM Application Ready Solution supports IBM Flex System x240 (x86) compute nodes.

- IBM System x helps organizations address their most challenging and complex problems. The Application Ready Solution supports IBM NeXtScale System, a revolutionary new x86 high-performance system designed for modular flexibility and scalability, System x rack-mounted servers and System x iDataPlex dx360 M4 systems designed to optimize density, performance and graphics acceleration for remote 3-D visualization.

- IBM System Storage Storwize V3700 is an entry-level disk system delivering an ideal price/performance ratio and scalability – or choose the optional IBM Storwize V7000 Unified for enterprise-class, midrange storage designed to consolidate block-and-file workloads into a single system.

- IBM Intelligent Cluster is a factory-integrated, fully tested solution that helps simplify and expedite deployment of x86-based Application Ready Solutions.

The solutions also include pre-integrated IBM Platform Computing software designed to address technical computing challenges:

- IBM Platform HPC is a complete technical computing management solution in a single product, with a range of features designed to improve time-to-results and help researchers focus on their work rather than on managing workloads.

- IBM Platform LSF provides a comprehensive set of tools for intelligently scheduling workloads and dynamically allocating resources to help ensure optimal job throughput.

- IBM Platform Symphony delivers powerful enterprise-class management for running big data, analytics and compute-intensive applications.

- IBM Platform Cluster Manager–Standard Edition provides easy-to-use yet powerful cluster management for technical computing clusters that simplifies the entire process, from initial deployment through provisioning to ongoing maintenance.

- IBM General Parallel File System (GPFS) is a high-performance enterprise file management platform for optimizing data management.
Technical computing workloads optimized for Application Ready Solutions
IBM Application Ready Solutions take the guesswork and complexity out of deploying, managing and using high-performance clusters, grids and clouds in industries such as automotive, aerospace, life sciences, electronics, telecommunications, chemistry and petroleum.

IBM Application Ready Solution for Abaqus
Developed in partnership with Dassault Systèmes, the IBM Application Ready Solution for Abaqus provides the framework to consolidate numerous tools into a single, unified modeling and analysis computing environment. High performance IBM systems, workload and file management, networking and storage combine to provide a complete integrated environment. Easy access to Abaqus job-related data and remote job management provides the means to solve entry-level or extremely large simulation problems with fast turnaround times.

IBM Application Ready Solution for Accelrys
Designed for healthcare and life sciences, the Application Ready Solution for Accelrys simplifies and accelerates mapping, variant calling and annotation for the Accelrys Enterprise Platform (AEP) NGS Collection. It addresses file system performance—the number-one challenge for NGS workloads on AEP—by integrating IBM GPFS for scalable I/O performance. IBM systems provide the computational power and high-performance storage required, along with simplified cluster management to speed deployment and provisioning.
IBM Application Ready Solution for ANSYS
ANSYS software helps engineers tackle demanding tasks such as computational fluid dynamics (CFD) modeling, structural analysis and digital wind-tunnel simulation. The IBM Application Ready Solution for ANSYS speeds deployment and optimizes performance for the most demanding ANSYS Fluent and ANSYS Mechanical environments. Engineers can become productive quickly, easily submitting simulations, sharing files with colleagues and enhancing insight when optional remote 2-D and 3-D visualization is configured.

IBM Application Ready Solution for CLC bio
This integrated solution is architected for clients involved in genomics research in areas ranging from personalized medicine to plant and food research. Combining CLC bio software with high-performance IBM systems and GPFS, the solution accelerates high-throughput sequencing and analysis of next-generation sequencing data while improving the efficiency of CLC bio Genomic Server and CLC Genomics Workbench environments.

IBM Application Ready Solution for Gaussian
Gaussian software is widely used by chemists, chemical engineers, biochemists, physicists and other scientists performing molecular electronic structure calculations in a variety of market segments. The IBM Application Ready Solution is designed to help speed results by integrating the latest version of the Gaussian series of programs with powerful IBM Flex System blades and integrated storage. IBM Platform Computing provides simplified workload and resource management.

IBM Application Ready Solution for IBM InfoSphere BigInsights
The Application Ready Solution for IBM InfoSphere BigInsights provides a powerful big data MapReduce analytics environment and reference architecture based on IBM PowerLinux servers, IBM Platform Symphony, IBM GPFS and integrated storage. The solution delivers balanced performance for data-intensive workloads, along with tools and accelerators to simplify and speed application development. The solution is ideal for solving time-critical, data-intensive analytics problems in a wide range of industry sectors.

IBM Application Ready Solution for MSC Software
The IBM Application Ready Solution for MSC Software features an optimized platform designed to help manufacturers rapidly deploy a high-performance simulation, modeling and data management environment, complete with process workflow and other high-demand usability features. The platform features IBM systems, workload management and parallel file system seamlessly integrated with MSC Nastran, MSC Patran and MSC SimManager to provide clients robust and agile engineering clusters or HPC clouds for accelerated results and lower cost.

IBM Application Ready Solution for Schlumberger
Fine-tuned for accelerating reservoir simulations using Schlumberger ECLIPSE and INTERSECT, this Application Ready Solution provides application templates to reduce setup time and simplify job submission. Architected specifically for Schlumberger applications, the solution enables users to perform significantly more iterations of their simulations and analysis, ultimately yielding more accurate results. Easy access to Schlumberger job-
related data and remote management improves user and administrator productivity.

**Complete, integrated solutions architected to deliver real-world benefits**

IBM Application Ready Solutions help organizations transform environments to deliver results faster, better and at less expense. The benefits start with pre-integrated and fully supported solutions that reduce the complexity of the IT lifecycle and shorten implementation time. Companies have one support number to call for all IBM software and hardware components for dedicated assistance from technical computing industry experts. Extensible IBM Application Ready Solutions also help protect a company’s technical computing investment by scaling as requirements grow. IBM Platform Computing software allows companies to speed time-to-results and lower costs by simplifying management of high-performance clusters and clouds. These products enable research and development teams to easily access a pool of shared resources to dramatically accelerate a wide range of simulations and analytics. Job submission templates reduce
setup time while minimizing user errors during job submission, and built-in workload management capabilities such as tracking application license usage and scheduling jobs based on license availability improve resource utilization and help ensure fastest time-to-results. Designed, tested and optimized by experienced technical computing architects from IBM and leading independent software vendors (ISVs), IBM Application Ready Solutions help deliver optimal application performance and robustness. IBM high-performance systems, software and storage are designed to accelerate even the most demanding workloads. Further performance improvements are provided by the advanced cluster file system, which improves efficiency and speed by removing data-related bottlenecks.
Enterprise-Ready Cluster & Workload Management

High performance computing (HPC) is becoming a necessary tool for organizations to speed up product design, scientific research, and business analytics. However, there are few software environments more complex to manage and utilize than modern high performance computing clusters. Therefore, addressing the problem of complexity in cluster management is a key aspect of leveraging HPC to improve time to results and user productivity.
Introduction

Clusters based on the Linux operating system have become increasingly prevalent at large supercomputing centers and continue to make significant in-roads in commercial and academic settings. This is primarily due to their superior price/performance and flexibility, as well as the availability of commercial applications that are based on the Linux OS.

Ironically, the same factors that make Linux a clear choice for high performance computing often make the operating system less accessible to smaller computing centers. These organizations may have Microsoft Windows administrators on staff, but have little or no Linux or cluster management experience. The complexity and cost of cluster management often outweigh the benefits that make open, commodity clusters so compelling. Not only can HPC cluster deployments be difficult, but the ongoing need to deal with heterogeneous hardware and operating systems, mixed workloads, and rapidly evolving toolsets make deploying and managing an HPC cluster a daunting task.

These issues create a barrier to entry for scientists and researchers who require the performance of an HPC cluster, but are limited to the performance of a workstation. This is why ease of use is now mandatory for HPC cluster management. This paper reviews the most complete and easy to use cluster management solution, Platform HPC, which is now commercially available from Platform Computing.

The cluster management challenge

To provide a proper HPC application environment, system administrators need to provide a full set of capabilities to their users, as shown below. These capabilities include cluster provisioning and node management, application workload management, and
an environment that makes it easy to develop, run and manage distributed parallel applications.

Modern application environments tend to be heterogeneous; some workloads require Windows compute hosts while others require particular Linux operating systems or versions. The ability to change a node’s operating system on-the-fly in response to changing application needs - referred to as adaptive scheduling - is important since it allows system administrators to maximize resource use, and present what appears to be a larger resource pool to cluster users.

Learning how to use a command line interface to power-up, provision and manage a cluster is extremely time-consuming. Administrators therefore need remote, web-based access to their HPC environment that makes it easier for them to install and manage an HPC cluster. An easy-to-use application-centric web interface can have tangible benefits including improved productivity, reduced training requirements, reduced errors rates, and secure remote access.

While there are several cluster management tools that address parts of these requirements, few address them fully, and some tools are little more than collections of discrete open-source software components.

Some cluster toolkits focus largely on the problem of cluster provisioning and management. While they clearly simplify cluster deployment, administrators wanting to make changes to node configurations or customize their environment will quickly find themselves hand-editing XML configuration files or writing their own shell scripts. Third-party workload managers and various open-source MPI libraries might be included as part of a distribution. However, these included components are loosely integrated and often need to be managed separately from the cluster manager. As a result the cluster administrator needs to learn how to utilize each additional piece of software in order to manage the cluster effectively.

Other HPC solutions are designed purely for application workload management. While these are all capable workload managers, most do not address at all the issue of cluster management, application integration, or adaptive scheduling. If such capabilities exist they usually require the purchase of additional software products.

Parallel job management is also critical. One of the primary reasons that customers deploy HPC clusters is to maximize application performance. Processing problems in parallel is a common way to achieve performance gains. The choice of MPI, its scalability, and the degree to which it is integrated with various OFED drivers and high performance interconnects has a direct impact on delivered application performance. Furthermore, if the workload manager does not incorporate specific parallel job management features, busy cluster users and administrators can find themselves manually cleaning up after failed MPI jobs or writing their own shell scripts to do the same.
Complexity is a real problem. Many small organizations or departments grapple with a new vocabulary full of cryptic commands, configuring and troubleshooting Anaconda kick start scripts, finding the correct OFED drivers for specialized hardware, and configuring open source monitoring systems like Ganglia or Nagios. Without an integrated solution administrators may need to deal with dozens of distinct software components, making managing HPC cluster implementations extremely tedious and time-consuming.

Re-thinking HPC clusters

Clearly these challenges demand a fresh approach to HPC cluster management. Platform HPC represents a “re-think” of how HPC clusters are deployed and managed. Rather than addressing only part of the HPC management puzzle, Platform HPC addresses all facets of cluster management. It provides:

- A complete, easy-to-use cluster management solution
- Integrated application support
- User-friendly, topology-aware workload management
- Robust workload and system monitoring and reporting
- Dynamic operating system multi-boot (adaptive scheduling)
- GPU scheduling
- Robust commercial MPI library (Platform MPI)
- Web-based interface for access anywhere

Most complete HPC cluster management solution

Platform HPC makes it easy to deploy, run and manage HPC clusters while meeting the most demanding requirements for application performance and predictable workload management. It is a complete solution that provides a robust set of cluster management capabilities; from cluster provisioning and management to workload
management and monitoring. The easy-to-use unified web portal provides a single point of access into the cluster, making it easy to manage your jobs and optimize application performance. Platform HPC is more than just a stack of software; it is a fully integrated and certified solution designed to ensure ease of use and simplified troubleshooting.

**Integrated application support**
High performing, HPC-optimized MPI libraries come integrated with Platform HPC, making it easy to get parallel applications up and running. Scripting guidelines and job submission templates for commonly used commercial applications simplify job submission, reduce setup time and minimize operation errors. Once the applications are up and running, Platform HPC improves application performance by intelligently scheduling resources based on workload characteristics.

**Fully certified and supported**
Platform HPC unlocks cluster management to provide the easiest and most complete HPC management capabilities while reducing overall cluster cost and improving administrator productivity. It is based on the industry’s most mature and robust workload manager, Platform LSF, making it the most reliable solution on the market. Other solutions are typically a collection of open-source tools, which may also include pieces of commercially developed software. They lack key HPC functionality and vendor support, relying on the administrator’s technical ability and time to implement. Platform HPC is a single product with a single installer and a unified web-based management interface. With the best support in the HPC industry, Platform HPC provides the most complete solution for HPC cluster management.

**Complete solution**
Platform HPC provides a complete set of HPC cluster management features. In this section we’ll explore some of these unique capabilities in more detail.

**Easy to use, cluster provisioning and management**
With Platform HPC, administrators can quickly provision and manage HPC clusters with unprecedented ease. It ensures maximum uptime and can transparently synchronize files to cluster nodes without any downtime or re-installation.

**Fast and efficient software Installation** – Platform HPC can be installed on the head node and takes less than one hour using three different mechanisms:
- Platform HPC DVD
- Platform HPC ISO file
- Platform partner’s factory install bootable USB drive
Installing software on cluster nodes is simply a matter of associating cluster nodes with flexible provisioning templates through the web-based interface.

**Flexible provisioning** – Platform HPC offers multiple options for provisioning Linux operating environments that include:
- Package-based provisioning
- Image based provisioning
- Diskless node provisioning

Large collections of hosts can be provisioned using the same provisioning template. Platform HPC automatically manages details such as IP address assignment and node naming conventions that reflect the position of cluster nodes in data center racks.
Unlike competing solutions, Platform HPC deploys multiple operating systems and OS versions to a cluster simultaneously. This includes Red Hat Enterprise Linux, CentOS, Scientific Linux, and SUSE Linux Enterprise Server. This provides administrators with greater flexibility in how they serve their user communities and means that HPC clusters can grow and evolve incrementally as requirements change.
What’s New in IBM Platform LSF 8

Written with Platform LSF administrators in mind, this brief provides a short explanation of significant changes in Platform’s latest release of Platform LSF, with a specific emphasis on scheduling and workload management features.

About IBM Platform LSF 8

Platform LSF is the most powerful workload manager for demanding, distributed high performance computing environments. It provides a complete set of workload management capabilities, all designed to work together to reduce cycle times and maximize productivity in mission-critical environments.

This latest Platform LSF release delivers improvements in performance and scalability while introducing new features that simplify administration and boost user productivity. This includes:

- **Guaranteed resources** – Aligns business SLA’s with infrastructure configuration for simplified administration and configuration
- **Live reconfiguration** – Provides simplified administration and enables agility
- **Delegation of administrative rights** – Empowers line of business owners to take control of their own projects
- **Fairshare & pre-emptive scheduling enhancements** – Fine tunes key production policies

Platform LSF 8 Features

Guaranteed Resources Ensure Deadlines are Met

In Platform LSF 8, resource-based scheduling has been extended to guarantee resource availability to groups of jobs. Resources can be slots, entire hosts or user-defined shared resources such as software licenses. As an example, a business unit might guarantee that it has access to specific types of resources within ten minutes of a job being submitted, even while sharing resources between departments. This facility ensures that lower priority jobs using the needed resources can be pre-empted in order to meet the SLAs of higher priority jobs.

Because jobs can be automatically attached to an SLA class via access controls, administrators can enable these guarantees without requiring that end-users change their job submission procedures, making it easy to implement this capability in existing environments.

Live Cluster Reconfiguration

Platform LSF 8 incorporates a new live reconfiguration capability, allowing changes to be made to clusters without the need to re-start LSF daemons. This is useful to customers who need to add hosts, adjust sharing policies or re-assign users between groups “on the fly”, without impacting cluster availability or running jobs.
Changes to the cluster configuration can be made via the `bconf` command line utility, or via new API calls. This functionality can also be integrated via a web-based interface using Platform Application Center. All configuration modifications are logged for a complete audit history, and changes are propagated almost instantaneously. The majority of reconfiguration operations are completed in under half a second.

With Live Reconfiguration, down-time is reduced, and administrators are free to make needed adjustments quickly rather than wait for scheduled maintenance periods or non-peak hours. In cases where users are members of multiple groups, controls can be put in place so that a group administrator can only control jobs associated with their designated group rather than impacting jobs related to another group submitted by the same user.

**Delegation of Administrative Rights**

With Platform LSF 8, the concept of group administrators has been extended to enable project managers and line of business managers to dynamically modify group membership and fairshare resource allocation policies within their group. The ability to make these changes dynamically to a running cluster is made possible by the Live Reconfiguration feature.

These capabilities can be delegated selectively depending on the group and site policy. Different group administrators can manage jobs, control sharing policies or adjust group membership.

**More Flexible Fairshare Scheduling Policies**

To enable better resource sharing flexibility with Platform LSF 8, the algorithms used to tune dynamically calculated user priorities can be adjusted at the queue level. These algorithms can vary based on
department, application or project team preferences. The Fairshare parameters ENABLE_HIST_RUN_TIME and HIST_HOURS enable administrators to control the degree to which LSF considers prior resource usage when determining user priority. The flexibility of Platform LSF 8 has also been improved by allowing a similar “decay rate” to apply to currently running jobs (RUN_TIME_DECAY), either system-wide or at the queue level. This is most useful for customers with long-running jobs, where setting this parameter results in a more accurate view of real resource use for the fairshare scheduling to consider.

Performance & Scalability Enhancements
Platform LSF has been extended to support an unparalleled scale of up to 100,000 cores and 1.5 million queued jobs for very high throughput EDA workloads. Even higher scalability is possible for more traditional HPC workloads. Specific areas of improvement include the time required to start the master-batch daemon (MBD), bjobs query performance, job submission and job dispatching as well as impressive performance gains resulting from the new Bulk Job Submission feature. In addition, on very large clusters with large numbers of user groups employing fairshare scheduling, the memory footprint of the master batch scheduler in LSF has been reduced by approximately 70% and scheduler cycle time has been reduced by 25%, resulting in better performance and scalability.

More Sophisticated Host-based Resource Usage for Parallel Jobs
Platform LSF 8 provides several improvements to how resource use is tracked and reported with parallel jobs. Accurate tracking of how parallel jobs use resources such as CPUs, memory and swap, is important for ease of management, optimal scheduling and accurate reporting and workload analysis. With Platform LSF 8 administrators can track resource usage on a per-host basis and an aggregated basis (across all hosts), ensuring that resource use is reported accurately. Additional details such as running PIDs and PGIDs for distributed parallel jobs, manual cleanup (if necessary) and the development of scripts for managing parallel jobs are simplified. These improvements in resource usage reporting are reflected in LSF commands including bjobs, bhist and bacct.

Improved Ease of Administration for Mixed Windows and Linux Clusters
The Ispasswd command in Platform LSF enables Windows LSF users to advise LSF of changes to their Windows level passwords. With Platform LSF 8, password synchronization between environments has become much easier to manage because the Windows passwords can now be adjusted directly from Linux hosts using the Ispasswd command. This allows Linux users to conveniently synchronize passwords on Windows hosts without needing to explicitly login into the host.

Bulk Job Submission
When submitting large numbers of jobs with different resource requirements or job level settings, Bulk Job Submission allows for jobs to be submitted in bulk by referencing a single file containing job details.
Enterprise-Ready Cluster & Workload Management
IBM Platform HPC

Simplified configuration changes – Platform HPC simplifies administration and increases cluster availability by allowing changes such as new package installations, patch updates, and changes to configuration files to be propagated to cluster nodes automatically without the need to re-install those nodes. It also provides a mechanism whereby experienced administrators can quickly perform operations in parallel across multiple cluster nodes.

Repository snapshots / trial installations – Upgrading software can be risky, particularly in complex environments. If a new software upgrade introduces problems, administrators often need to rapidly “rollback” to a known good state. With other cluster managers this can mean having to re-install the entire cluster. Platform HPC incorporates repository snapshots, which are “restore points” for the entire cluster. Administrators can snapshot a known good repository, make changes to their environment, and easily revert to a previous “known good” repository in the event of an unforeseen problem. This powerful capability takes the risk out of cluster software upgrades.

New hardware integration – When new hardware is added to a cluster it may require new or updated device drivers that are not supported by the OS environment on the installer node. This means that a newly updated node may not network boot and provision until the head node on the cluster is updated with a new operating system; a tedious and disruptive process. Platform HPC includes a driver patching utility that allows updated device drivers to be inserted into existing repositories, essentially future proofing the cluster, and providing a simplified means of supporting new hardware without needing to re-install the environment from scratch.
**Software updates with no re-boot** – Some cluster managers always re-boot nodes when updating software, regardless of how minor the change. This is a simple way to manage updates. However, scheduling downtime can be difficult and disruptive. Platform HPC performs updates intelligently and selectively so that compute nodes continue to run even as non-intrusive updates are applied. The repository is automatically updated so that future installations include the software update. Changes that require the re-installation of the node (e.g. upgrading an operating system) can be made in a “pending” state until downtime can be scheduled.

**User-friendly, topology aware workload management**

Platform HPC includes a robust workload scheduling capability, which is based on Platform LSF - the industry's most powerful, comprehensive, policy driven workload management solution for engineering and scientific distributed computing environments. By scheduling workloads intelligently according to policy, Platform HPC improves end user productivity with minimal system administrative effort. In addition, it allows HPC user teams to easily access and share all computing resources, while reducing time between simulation iterations.

**GPU scheduling** – Platform HPC provides the capability to schedule jobs to GPUs as well as CPUs. This is particularly advantageous in heterogeneous hardware environments as it means that administrators can configure Platform HPC so that only those jobs that can benefit from running on GPUs are allocated to those resources. This frees up CPU-based resources to run other jobs. Using the unified management interface, administrators can monitor the GPU performance as well as detect ECC errors.

**Unified management interface**

Competing cluster management tools either do not have a web-based interface or require multiple interfaces for managing different functional areas. In comparison, Platform HPC includes a single unified interface through which all administrative tasks can be performed including node-management, job-management, jobs and cluster monitoring and reporting. Using the unified management interface, even cluster administrators with very little Linux experience can competently manage a state of the art HPC cluster.

**Job management** – While command line savvy users can continue using the remote terminal capability, the unified web portal makes it easy to submit, monitor, and manage jobs. As changes are made to the cluster configuration, Platform HPC automatically re-configures key components, ensuring that jobs are allocated to the appropriate resources.

The web portal is customizable and provides job data management, remote visualization and interactive job support.
**Workload/system correlation** – Administrators can correlate workload information with system load, so that they can make timely decisions and proactively manage compute resources against business demand. When it’s time for capacity planning, the management interface can be used to run detailed reports and analyses which quantify user needs and remove the guess work from capacity expansion.

**Simplified cluster management** – The unified management console is used to administer all aspects of the cluster environment. It enables administrators to easily install, manage and monitor their cluster. It also provides an interactive environment to easily package software as kits for application deployment as well as pre-integrated commercial application support. One of the key features of the interface is an operational dashboard that provides comprehensive administrative reports. As the image illustrates, Platform HPC enables administrators to monitor and report on key performance metrics such as cluster capacity, available memory and CPU utilization. This enables administrators to easily identify and troubleshoot issues.

The easy to use interface saves the cluster administrator time, and means that they do not need to become an expert in the administration of open-source software components. It also reduces the possibility of errors and time lost due to incorrect configuration. Cluster administrators enjoy the best of both worlds – easy access to a powerful, web-based cluster manager without the need to learn and separately administer all the tools that comprise the HPC cluster environment.
Robust Commercial MPI library
Platform MPI – In order to make it easier to get parallel applications up and running, Platform HPC includes the industry’s most robust and highest performing MPI implementation, Platform MPI. Platform MPI provides consistent performance at application run-time and for application scaling, resulting in top performance results across a range of third-party benchmarks.

Open Source MPI – Platform HPC also includes various other industry standard MPI implementations. This includes MPICH1, MPICH2 and MVAPICH1, which are optimized for cluster hosts connected via InfiniBand, iWARP or other RDMA based interconnects.

Integrated application support
Job submission templates – Platform HPC comes complete with job submission templates for ANSYS Mechanical, ANSYS Fluent, ANSYS CFX, LS-DYNA, MSC Nastran, Schlumberger ECLIPSE, Simulia Abaqus, NCBI Blast, NWChem, ClustalW, and HMMER. By configuring these templates based on the application settings in your environment, users can start using the cluster without writing scripts.

Scripting Guidelines – Cluster users that utilize homegrown or open-source applications, can utilize the Platform HPC scripting guidelines. These user-friendly interfaces help minimize job submission errors. They are also self-documenting, enabling users to create their own job submission templates.

Benchmark tests – Platform HPC also includes standard benchmark tests to ensure that your cluster will deliver the best performance without manual tuning.

Flexible OS provisioning
Platform HPC can deploy multiple operating system versions concurrently on the same cluster and, based on job resource requirements, dynamically boot the Linux or Windows operating system required to run the job. Administrators can also use a web interface to manually switch nodes to the required OS to meet application demands, providing them with the flexibility to support special requests and accommodate unanticipated changes. Rather than being an extracost item as it is with other HPC management suites, this capability is included as a core feature of Platform HPC.

Commercial Service and support
Certified cluster configurations – Platform HPC is tested and certified on all partner hardware platforms. By qualifying each platform individually and providing vendor-specific software with optimized libraries and drivers that take maximum advantage of unique hardware features, Platform Computing has essentially done the integration work in advance.
As a result, clusters can be deployed quickly and predictably with minimal effort. As a testament to this, Platform HPC is certified under the Intel Cluster Ready program.

**Enterprise class service and support** – Widely regarded as having the best HPC support organization in the business, Platform Computing is uniquely able to support an integrated HPC platform. Because support personnel have direct access to the Platform HPC developers, Platform Computing is able to offer a higher level of support and ensure that any problems encountered are resolved quickly and efficiently.

**Summary**
Platform HPC is the ideal solution for deploying and managing state-of-the-art HPC clusters. It makes cluster management simple, enabling analysts, engineers, and scientists from organizations of any size to easily exploit the power of Linux clusters. Unlike other HPC solutions that address only parts of the HPC management challenge, Platform HPC uniquely addresses all aspects of cluster and management including:

- Easy-to-use cluster provisioning and management
- User-friendly, topology aware workload management
- Unified management interface
- Robust commercial MPI library
- Integrated application support
- Flexible OS provisioning
- Commercial HPC service and support
By providing simplified management over the entire lifecycle of a cluster, Platform HPC has a direct and positive impact on productivity while helping to reduce complexity and cost. The comprehensive web-based management interface, and features like repository snapshots and the ability to update software packages on the fly means that state-of-the-art HPC clusters can be provisioned and managed even by administrators with little or no Linux administration experience.

<table>
<thead>
<tr>
<th>Capability / Feature</th>
<th>Platform HPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster Provisioning and Management</td>
<td>✓</td>
</tr>
<tr>
<td>Initial cluster provisioning</td>
<td>✓</td>
</tr>
<tr>
<td>Multiple provisioning methods</td>
<td>✓</td>
</tr>
<tr>
<td>Web-based cluster management</td>
<td>✓</td>
</tr>
<tr>
<td>Node updates with no re-boot</td>
<td>✓</td>
</tr>
<tr>
<td>Repository snapshots</td>
<td>✓</td>
</tr>
<tr>
<td>Flexible node templates</td>
<td>✓</td>
</tr>
<tr>
<td>Multiple OS and OS versions</td>
<td>✓</td>
</tr>
<tr>
<td>Workload Management &amp; Application Integration</td>
<td>✓</td>
</tr>
<tr>
<td>Integrated workload management</td>
<td>✓</td>
</tr>
<tr>
<td>HPC libraries &amp; toolsets</td>
<td>✓</td>
</tr>
<tr>
<td>NVIDIA CUDA SDK support</td>
<td>✓</td>
</tr>
<tr>
<td>Web-based job management</td>
<td>✓</td>
</tr>
<tr>
<td>Web-based job data management</td>
<td>✓</td>
</tr>
<tr>
<td>Multi-boot based on workload</td>
<td>✓</td>
</tr>
<tr>
<td>Advanced parallel job management</td>
<td>✓</td>
</tr>
<tr>
<td>Commercial application integrations</td>
<td>✓</td>
</tr>
<tr>
<td>MPI Libraries</td>
<td>✓</td>
</tr>
<tr>
<td>Commercial grade MPI</td>
<td>✓</td>
</tr>
<tr>
<td>Workload and system monitoring, reporting and correlation</td>
<td>✓</td>
</tr>
<tr>
<td>Workload monitoring</td>
<td>✓</td>
</tr>
<tr>
<td>Workload reporting</td>
<td>✓</td>
</tr>
<tr>
<td>System monitoring &amp; reporting</td>
<td>✓</td>
</tr>
<tr>
<td>Workload and system load correlation</td>
<td>✓</td>
</tr>
<tr>
<td>Integration with 3rd party management tools</td>
<td>✓</td>
</tr>
</tbody>
</table>
IBM Platform MPI 8.1

Benefits
- Superior application performance
- Reduced development and support costs
- Faster time-to-market
- The industry's best technical support

Features
- Supports the widest range of hardware, networks and operating systems
- Distributed by over 30 leading commercial software vendors
- Change interconnects or libraries with no need to re-compile
- Seamless compatibility across Windows and Linux environments
- Ensures a production quality implementation

Ideal for:
- Enterprises that develop or deploy parallelized software applications on HPC clusters
- Commercial software vendors wanting to improve applications performance over the widest range of computer hardware, interconnects and operating systems

The Standard for Scalable, Parallel Applications
Platform MPI is a high performance, production-quality implementation of the Message Passing Interface (MPI). It is widely used in the high performance computing (HPC) industry and is considered the de facto standard for developing scalable, parallel applications.
Platform MPI maintains full backward compatibility with HP-MPI and Platform MPI applications and incorporates advanced CPU affinity features, dynamic selection of interface libraries, superior workload manager integrations and improved performance and scalability.
Platform MPI supports the broadest range of industry standard platforms, interconnects and operating systems helping ensure that your parallel applications can run anywhere.

**Focus on portability**
Platform MPI allows developers to build a single executable that transparently leverages the performance features of any type of interconnect, thereby providing applications with optimal latency and bandwidth for each protocol. This reduces development effort, and enables applications to use the “latest and greatest” technologies on Linux or Microsoft Windows without the need to re-compile and re-link applications.

Platform MPI is optimized for both distributed (DMP) and shared memory (SMP) environments and provides a variety of flexible CPU binding strategies for processes and threads, enabling better performance on multi-core environments. With this capability memory and cache conflicts are managed by more intelligently distributing the load among multiple cores.
With support for Windows HPC Server 2008 and the Microsoft job scheduler, as well as other Microsoft operating environments, Platform MPI allows developers targeting Windows platforms to enjoy the benefits of a standard portable MPI and avoid proprietary lock-in.
Enterprise-Ready Cluster & Workload Management
IBM Platform MPI 8.1

**Supported Operating Systems**
- Red Hat Enterprise Linux 4.6, 5.x and 6.x
- SUSE Linux Enterprise Server 10 and 11
- CentOS 5.3

**Supported Interconnects and Protocols**
- **Myrinet (Linux)**: GM & MX on X86-64 and Itanium2
- **InfiniBand (Linux)**: OFED, PSM, uDAPL on X86-64 and Itanium2
  - OFED 1.1, 1.2, 1.3, 1.4, 1.5
  - SDR, DDR, QDR, ConnectX and ConnextX2 Mellanox FCA
- **GigE (Linux)**: RDMA, uDAPL, TCP/IP
- **Infiniband (Windows)**: WinOF 2.x, IBAL, WSD, SDR, DDR, QDR, ConnectX(2)
- **GigE (Windows)**: TCP/IP on x86-64
## Features and Benefits

<table>
<thead>
<tr>
<th>Features</th>
<th>Benefits</th>
</tr>
</thead>
</table>
| **Simplicity** | - Fully complies with the MPI 2.2 standard, providing dynamic processes, one-sided communications, extended collectives, thread safety, and updated ROMIO  
- Complete debugging, diagnostic and profiling tools  
- Auto-detection of interconnects and dynamic loading of libraries  
- No re-link required for debugging and profiling  
- Supported by the largest dedicated HPC support organization |
| **Performance** | - Applications port easily to other platforms  
- Protects ISV software investment  
- Reduces time-to-market  
- Increased robustness and quality of applications  
- Technical problems resolved quickly and efficiently |
| **Compatibility** | - Takes maximum advantage of available hardware  
- Reduced latency for better performance  
- Performance improves without explicit developer action  
- Better message throughput in streaming applications  
- Easier to optimize application performance |
| **Flexibility** | - Avoid the cost of separate releases for different platforms  
- Easily used with existing MPI applications  
- Common mpirun syntax between Linux and Windows  
- Customers avoid proprietary “lock-in”  
- Avoid floating point issues causing inconsistent results |
| **Flexibility** | - Supports the widest variety of networks and interconnects  
- Select interconnects at run-time with no need to re-compile  
- Write applications once and deploy across multiple OS and hardware topologies  
- CPU binding features well suited to GPU-aware applications |
| **Compatibility** | - Develop applications that will run on more platforms  
- Reduce testing, maintenance and support costs  
- Enjoy strategic flexibility |
**transtec HPC as a Service**
You will get a range of applications like LS-Dyna, ANSYS, Gromacs, NAMD etc. from all kinds of areas pre-installed, integrated into an enterprise-ready cloud and workload management system, and ready-to run. Do you miss your application?

*Ask us: HPC@transtec.de*

**transtec Platform as a Service**
You will be provided with dynamically provided compute nodes for running your individual code. The operating system will be pre-installed according to your requirements. Common Linux distributions like RedHat, CentOS, or SLES are the standard. Do you need another distribution?

*Ask us: HPC@transtec.de*

**transtec Hosting as a Service**
You will be provided with hosting space inside a professionally managed and secured datacenter where you can have your machines hosted, managed, maintained, according to your requirements. Thus, you can build up your own private cloud. What range of hosting and maintenance services do you need?

*Tell us: HPC@transtec.de*

---

**HPC @ transtec: Services and Customer Care from A to Z**

transtec AG has over 30 years of experience in scientific computing and is one of the earliest manufacturers of HPC clusters. For nearly a decade, transtec has delivered highly customized High Performance clusters based on standard components to academic and industry customers across Europe with all the high quality standards and the customer-centric approach that transtec is well known for.

Every transtec HPC solution is more than just a rack full of hardware – it is a comprehensive solution with everything the HPC user, owner, and operator need.

In the early stages of any customer’s HPC project, transtec experts provide extensive and detailed consulting to the customer – they benefit from expertise and experience. Consulting is followed by benchmarking of different systems with either specifically crafted

---

**Services and Customer Care from A to Z**

- individual Presales consulting
- benchmarking of different systems
- continual improvement
- maintenance, support & managed services
- application-, customer-, site-specific sizing of HPC solution
- burn-in tests of systems
- integration into customer’s environment
- onsite hardware assembly
- application installation
- software & OS installation
- customer training
customer code or generally accepted benchmarking routines; this aids customers in sizing and devising the optimal and detailed HPC configuration.

Each and every piece of HPC hardware that leaves our factory undergoes a burn-in procedure of 24 hours or more if necessary. We make sure that any hardware shipped meets our and our customers’ quality requirements. transtec HPC solutions are turnkey solutions. By default, a transtec HPC cluster has everything installed and configured – from hardware and operating system to important middleware components like cluster management or developer tools and the customer’s production applications. Onsite delivery means onsite integration into the customer’s production environment, be it establishing network connectivity to the corporate network, or setting up software and configuration parts.

transtec HPC clusters are ready-to-run systems – we deliver, you turn the key, the system delivers high performance. Every HPC project entails transfer to production: IT operation processes and policies apply to the new HPC system. Effectively, IT personnel is trained hands-on, introduced to hardware components and software, with all operational aspects of configuration management.

transtec services do not stop when the implementation projects ends. Beyond transfer to production, transtec takes care: transtec offers a variety of support and service options, tailored to the customer’s needs. When you are in need of a new installation, a major reconfiguration or an update of your solution – transtec is able to support your staff and, if you lack the resources for maintaining the cluster yourself, maintain the HPC solution for you. From Professional Services to Managed Services for daily operations and required service levels, transtec will be your complete HPC service and solution provider. transtec’s high standards of performance, reliability and dependability assure your productivity and complete satisfaction.

transtec’s offerings of HPC Managed Services offer customers the possibility of having the complete management and administration of the HPC cluster managed by transtec service specialists, in an ITIL compliant way. Moreover, transtec’s HPC on Demand services help provide access to HPC resources whenever they need them, for example, because they do not have the possibility of owning and running an HPC cluster themselves, due to lacking infrastructure, know-how, or admin staff.

transtec HPC Cloud Services

Last but not least transtec’s services portfolio evolves as customers’ demands change. Starting this year, transtec is able to provide HPC Cloud Services. transtec uses a dedicated datacenter to provide computing power to customers who are in need of more capacity than they own, which is why this workflow model is sometimes called computing-on-demand. With these dynamically provided resources, customers with the possibility to have their jobs run on HPC nodes in a dedicated datacenter, professionally managed and secured, and individually customizable. Numerous standard applications like ANSYS, LS-Dyna, OpenFOAM, as well as lots of codes like Gromacs, NAMD, VMD, and others are pre-installed, integrated into an enterprise-ready cloud and workload management environment, and ready to run.

Alternatively, whenever customers are in need of space for hosting their own HPC equipment because they do not have the space capacity or cooling and power infrastructure themselves, transtec is also able to provide Hosting Services to those customers who’d like to have their equipment professionally hosted, maintained, and managed. Customers can thus build up their own private cloud!

Are you interested in any of transtec’s broad range of HPC related services? Write us an email to HPC@transtec.de. We’ll be happy to hear from you!
Scalable & Energy Efficient HPC Systems

There is no end in sight to growing data and computing requirements – which poses a serious challenge for space-constrained data centers. Also challenging for today’s organizations is the need to perform a larger number and variety of functions – without increasing budgets. IBM NeXtScale System, an economical addition to the IBM System x family, offers an innovative approach to maximum usable density.
Optimized to handle a number of workloads, all demanding agility, NeXtScale System helps drive business velocity by providing rapid procurement, deployment and flexible options. This simple, yet powerful, system can handle applications ranging from technical computing, to grid deployments, to analytics workloads, to large-scale cloud and virtualization infrastructures. Designed with industry-standard, off-the-shelf components, this general-purpose platform enables users to create a flexible, mix-and-match offering with compute, storage, and acceleration via graphics processing unit (GPU) or Intel Xeon Phi coprocessor. Customized solutions can be configured to provide application-appropriate platform with choice of servers, networking switches, adapters, and racks.

This modular system is designed to scale and grow along with data center needs in order to protect and maximize IT investments. Since it is optimized for standard racks, users can easily mix high-density NeXtScale server offerings and non-NeXtScale components within the same rack. NeXtScale System also provides tremendous time to value by enabling users to get it up and running – and to the production phase – faster.

**Building upon a strong System x foundation**

Extending the System x family to a larger range of users, the customizable, space-saving NeXtScale System comprises powerful compute nodes and an energy-efficient, low-cost 12-bay chassis.

**IBM NeXtScale nx360 M4 server**

This powerful server provides a dense, flexible solution with a low total cost of ownership. The half-wide, dual-socket NeXtScale nx360 M4 server is designed for data centers...
that require high performance but are constrained by floor space. By taking up less physical space in the data center, the NeXtScale server significantly enhances density. And it supports Intel Xeon E5-2600 v2 series up to 130 W and 12-core processors thus providing more performance per server. The nx360 M4 compute node contains only essential components in the base architecture to provide a cost-optimized platform.

Flexible, IT your way
Developed at the solution level, the NeXtScale System architecture is extremely flexible – enabling different technologies to easily fit into its design, for varied workloads. And since the system allows compute, storage, and acceleration via GPU or Intel Xeon Phi coprocessor to share the same chassis and architecture, it is very easy to deploy and grow. Front-access cabling – either from the bottom or the top of the rack – and direct-dock power capabilities enable users to make quick and easy changes to nodes, cables and networking switches. Plus, NeXtScale System supports multiple networking topologies, including Ethernet, InfiniBand and Fibre Channel.
System flexibility even extends to procurement: Organizations can either receive the system fully configured, pretested, IBM installed, and ready to power on; or self-configure and install using existing components to build a custom system.

**Simple yet elegant**

NeXtScale System makes choosing the right architecture for individual applications, budgets and data centers simple and economical. It optimizes shared infrastructure with common fans and power supplies leaving nodes to be completely independent and self-sufficient. The nodes do not share resources such as disks or memory. To manage costs, only essential components are included in the base architecture, and nodes can be used for either storage or GPU/coprocessor acceleration. This enables NeXtScale for an easy insertion into your infrastructure with your current tools and best practices. The ability of NeXtScale System to work with any standard switch, rack or networking card provides almost unlimited options to space- and budget-conscious organizations in even the most demanding industries.

**Scale for everyone**

The high-performance NeXtScale System enables organizations of all sizes and budgets to start small and scale rapidly, as needed, into future requirements. Rather than requiring organizations to purchase large clusters, this system offers a complete building-block approach in which users can start out with one chassis and add systems and components as needed. Designed to be easily run and simply managed at any scale – from a handful to thousands – NeXtScale System can help organizations achieve maximum impact per dollar.
<table>
<thead>
<tr>
<th><strong>IBM NeXtScale nx360 M4 at a glance</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Form factor/height</strong></td>
<td>Half-wide 1U</td>
</tr>
<tr>
<td><strong>Processor</strong></td>
<td>Two Intel Xeon E5-2600 v2 series</td>
</tr>
</tbody>
</table>
| **Cache**                    | Level 2: 256 KB per core  
| **Memory**                   | 8 DDR3/DDR3L LP, 128 GB maximum with 16 GB LP RDIMM |
| **Chassis support**          | NeXtScale n1200 Enclosure |
| **Local Storage**            | One 3.5-inch, two 2.5-inch SAS/SATA hard disk drives (HDDs) or four 1.8-inch solid state drives, up to 4 TB maximum capacity with one 4 TB 3.5-inch HDD |
| **Storage Native Expansion (NEX) Tray** | Eight 3.5-inch SAS/SATA HDDs, up to 32 TB maximum capacity |
| **Internal RAID**            | Onboard SATA controller with RAID options |
| **USB ports**                | One internal USB key |
| **Ethernet**                 | Two built-in 1 Gigabit Ethernet (GbE) ports standard |
| **Input/output**             | Two InfiniBand FDR ports (slotless option), two 10 GbE (slotless option), one PCIe (x16 PCI Express 3.0) |
| **Power management**         | Rack-level power capping and management via IBM Extreme Cloud Administration Toolkit (xCAT) |
| **Systems management**       | IBM Integrated Management Module 2 (IMM2) with dedicated management port, IPMI 2.0 compliant, Platform LSF and Platform HPC |
| **Operating systems supported** | Microsoft Windows Server, SUSE Linux Enterprise Server, Red Hat Enterprise Linux, VMware vSphere Hypervisor (ESXi) |
| **Limited warranty**         | 3-year customer replaceable unit and onsite limited warranty, next business day 9x5, service upgrades available |
Scalable, Energy Efficient
HPC Systems
IBM NeXtScale System
<table>
<thead>
<tr>
<th>IBM NeXtScale n1200 Enclosure at a glance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form factor</td>
</tr>
<tr>
<td>Bays</td>
</tr>
<tr>
<td>Power supply</td>
</tr>
<tr>
<td>Fans</td>
</tr>
<tr>
<td>Controller</td>
</tr>
</tbody>
</table>
Big Data, Cloud Storage it doesn’t matter what you call it, there is certainly increasing demand to store larger and larger amounts of unstructured data.

The IBM General Parallel File System (GPFS) has always been considered a pioneer of big data storage and continues today to lead in introducing industry leading storage technologies. Since 1998 GPFS has lead the industry with many technologies that make the storage of large quantities of file data possible. The latest version continues in that tradition, GPFS 3.5 represents a significant milestone in the evolution of big data management. GPFS 3.5 introduces some revolutionary new features that clearly demonstrate IBM’s commitment to providing industry leading storage solutions.
General Parallel File System (GPFS)

What is GPFS?
GPFS is more than clustered file system software; it is a full featured set of file management tools. This includes advanced storage virtualization, integrated high availability, automated tiered storage management and the performance to effectively manage very large quantities of file data.

GPFS allows a group of computers concurrent access to a common set of file data over a common SAN infrastructure, a network or a mix of connection types. The computers can run any mix of AIX, Linux or Windows Server operating systems. GPFS provides storage management, information life cycle management tools, centralized administration and allows for shared access to file systems from remote GPFS clusters providing a global namespace.

A GPFS cluster can be a single node, two nodes providing a high availability platform supporting a database application, for example, or thousands of nodes used for applications like the modeling of weather patterns. The largest existing configurations exceed 5,000 nodes. GPFS has been available since 1998 and has been field proven for more than 14 years on some of the world’s most powerful supercomputers to provide reliability and efficient use of infrastructure bandwidth.

GPFS was designed from the beginning to support high performance parallel workloads and has since been proven very effective for a variety of applications. Today it is installed in clusters supporting big data analytics, gene sequencing, digital media and scalable file serving. These applications are used across many industries including financial, retail, digital media, biotechnology, science and government. GPFS continues to push technology limits by being deployed in very demanding large environments. You may not need multiple petabytes of
data today, but you will, and when you get there you can rest assured GPFS has already been tested in these environments. This leadership is what makes GPFS a solid solution for any size application.

Supported operating systems for GPFS Version 3.5 include AIX, Red Hat, SUSE and Debian Linux distributions and Windows Server 2008.

The file system
A GPFS file system is built from a collection of arrays that contain the file system data and metadata. A file system can be built from a single disk or contain thousands of disks storing petabytes of data. Each file system can be accessible from all nodes within the cluster. There is no practical limit on the size of a file system. The architectural limit is 299 bytes. As an example, current GPFS customers are using single file systems up to 5.4PB in size and others have file systems containing billions of files.

Application interfaces
Applications access files through standard POSIX file system interfaces. Since all nodes see all of the file data applications can scale-out easily. Any node in the cluster can concurrently read or update a common set of files. GPFS maintains the coherency and consistency of the file system using sophisticated byte range locking, token (distributed lock) management and journaling. This means that applications using standard POSIX locking semantics do not need to be modified to run successfully on a GPFS file system.

In addition to standard interfaces GPFS provides a unique set of extended interfaces which can be used to provide advanced application functionality. Using these extended interfaces an application can determine the storage pool placement of a file, create a file clone and manage quotas. These extended interfaces provide features in addition to the standard POSIX interface.

Performance and scalability
GPFS provides unparalleled performance for unstructured data. GPFS achieves high performance I/O by:
- Striping data across multiple disks attached to multiple nodes.
- Supporting a wide range of file system block sizes to match I/O requirements.
- Utilizing advanced algorithms to improve read-ahead and write-behind IO operations.
- Using block level locking based on a very sophisticated scalable token management system to provide data consistency while allowing multiple application nodes concurrent access to the files.

When creating a GPFS file system you provide a list of raw devices and they are assigned to GPFS as Network Shared Disks (NSD). Once a NSD is defined all of the nodes in the GPFS cluster can access the disk, using local disk connection, or using the GPFS NSD network protocol for shipping data over a TCP/IP or InfiniBand connection. GPFS token (distributed lock) management coordinates access to NSD’s ensuring the consistency of file system data and metadata when different nodes access the same file. Token management responsibility is dynamically allocated among designated nodes in the cluster. GPFS can assign one or more nodes to act as token managers for a single file system. This allows greater scalability when you have a large number of files with high transaction workloads. In the event of a node failure the token management responsibility is moved to another node.
All data stored in a GPFS file system is striped across all of the disks within a storage pool, whether the pool contains 2 LUNS or 2,000 LUNS. This wide data striping allows you to get the best performance for the available storage. When disks are added to or removed from a storage pool existing file data can be redistributed across the new storage to improve performance. Data redistribution can be done automatically or can be scheduled. When redistributing data you can assign a single node to perform the task to control the impact on a production workload or have all of the nodes in the cluster participate in data movement to complete the operation as quickly as possible. Online storage configuration is a good example of an enterprise class storage management feature included in GPFS.

To achieve the highest possible data access performance GPFS recognizes typical access patterns including sequential, reverse sequential and random optimizing I/O access for these patterns.

Along with distributed token management, GPFS provides scalable metadata management by allowing all nodes of the cluster accessing the file system to perform file metadata operations. This feature distinguishes GPFS from other cluster file systems which typically have a centralized metadata server handling fixed regions of the file namespace. A centralized metadata server can often become a performance bottleneck for metadata intensive operations, limiting scalability and possibly introducing a single point of failure. GPFS solves this problem by enabling all nodes to manage metadata.

**Administration**

GPFS provides an administration model that is easy to use and is consistent with standard file system administration practices while providing extensions for the clustering aspects of GPFS. These functions support cluster management and other standard file system
administration functions such as user quotas, snapshots and extended access control lists.

GPFS administration tools simplify cluster-wide tasks. A single GPFS command can perform a file system function across the entire cluster and most can be issued from any node in the cluster. Optionally you can designate a group of administration nodes that can be used to perform all cluster administration tasks, or only authorize a single login session to perform admin commands cluster-wide. This allows for higher security by reducing the scope of node to node administrative access.

Rolling upgrades allow you to upgrade individual nodes in the cluster while the file system remains online. Rolling upgrades are supported between two major version levels of GPFS (and service levels within those releases). For example you can mix GPFS 3.4 nodes with GPFS 3.5 nodes while migrating between releases.

Quotas enable the administrator to manage file system usage by users and groups across the cluster. GPFS provides commands to generate quota reports by user, group and on a sub-tree of a file system called a fileset. Quotas can be set on the number of files (inodes) and the total size of the files. New in GPFS 3.5 you can now define user and group per fileset quotas which allows for more options in quota configuration. In addition to traditional quota management, the GPFS policy engine can be used query the file system metadata and generate customized space usage reports.

An SNMP interface allows monitoring by network management applications. The SNMP agent provides information on the GPFS cluster and generates traps when events occur in the cluster. For example, an event is generated when a file system is mounted or if a node fails. The SNMP agent runs on Linux and AIX. You can monitor a heterogeneous cluster as long as the agent runs on a Linux or AIX node.

You can customize the response to cluster events using GPFS callbacks. A callback is an administrator defined script that is executed when an event occurs, for example, when a file system is un-mounted or a file system is low on free space. Callbacks can be used to create custom responses to GPFS events and integrate these notifications into various cluster monitoring tools.

GPFS provides support for the Data Management API (DMAPI) interface which is IBM’s implementation of the X/Open data storage management API. This DMAPI interface allows vendors of storage management applications such as IBM Tivoli® Storage Manager (TSM) and High Performance Storage System (HPSS) to provide Hierarchical Storage Management (HSM) support for GPFS.

GPFS supports POSIX and NFS V4 access control lists (ACLs). NFS v4 ACLs can be used to serve files using NFSv4, but can also be used in other deployments, for example, to provide ACL support to nodes running Windows. To provide concurrent access from multiple operating system types GPFS allows you to run mixed POSIX and NFS v4 permissions in a single file system and map user and group IDs between Windows and Linux/UNIX environments.

File systems may be exported to clients outside the cluster through NFS. GPFS is often used as the base for a scalable NFS file service infrastructure. The GPFS clustered NFS (cNFS) feature provides data availability to NFS clients by providing NFS service continuation if an NFS server fails. This allows a GPFS cluster to provide scalable file service by providing simultaneous access to a common set of data from multiple nodes. The clustered NFS tools include monitoring of file services and IP address fail over. GPFS cNFS supports NFSv3 only. You can export a GPFS file system using NFSv4 but not with cNFS.
General Parallel File System (GPFS)

Technologies That Enable the Management of Big Data

Data availability

GPFS is fault tolerant and can be configured for continued access to data even if cluster nodes or storage systems fail. This is accomplished though robust clustering features and support for synchronous and asynchronous data replication.

GPFS software includes the infrastructure to handle data consistency and availability. This means that GPFS does not rely on external applications for cluster operations like node failover. The clustering support goes beyond who owns the data or who has access to the disks. In a GPFS cluster all nodes see all of the data and all cluster operations can be done by any node in the cluster with a server license. All nodes are capable of performing all tasks. What tasks a node can perform is determined by the type of license and the cluster configuration.

As a part of the built-in availability tools GPFS continuously monitors the health of the file system components. When failures are detected appropriate recovery action is taken automatically. Extensive journaling and recovery capabilities are provided which maintain metadata consistency when a node holding locks or performing administrative services fails.

Snapshots can be used to protect the file system’s contents against a user error by preserving a point in time version of the file system or a sub-tree of a file system called a fileset. GPFS implements a space efficient snapshot mechanism that generates a map of the file system or fileset at the time the snapshot is taken. New data blocks are consumed only when the file system data has been deleted or modified after the snapshot was created. This is done using a redirect-on-write technique (sometimes called copy-on-write). Snapshot data is placed in existing storage pools simplifying administration and optimizing the use of existing storage. The snapshot function can be used with a
backup program, for example, to run while the file system is in use and still obtain a consistent copy of the file system as it was when the snapshot was created. In addition, snapshots provide an online backup capability that allows files to be recovered easily from common problems such as accidental file deletion.

**Data Replication**

For an additional level of data availability and protection synchronous data replication is available for file system metadata and data. GPFS provides a very flexible replication model that allows you to replicate a file, set of files, or an entire file system. The replication status of a file can be changed using a command or by using the policy based management tools. Synchronous replication allows for continuous operation even if a path to an array, an array itself or an entire site fails.

Synchronous replication is location aware which allows you to optimize data access when the replicas are separated across a WAN. GPFS has knowledge of what copy of the data is “local” so read-heavy applications can get local data read performance even when data replicated over a WAN. Synchronous replication works well for many workloads by replicating data across storage arrays within a data center, within a campus or across geographical distances using high quality wide area network connections.

When wide area network connections are not high performance or are not reliable, an asynchronous approach to data replication is required. GPFS 3.5 introduces a feature called Active File Management (AFM). AFM is a distributed disk caching technology developed at IBM Research that allows the expansion of the GPFS global namespace across geographical distances. It can be used to provide high availability between sites or to provide local “copies” of data distributed to one or more GPFS clusters. For more details on AFM see the section entitled Sharing data between clusters.

For a higher level of cluster reliability GPFS includes advanced clustering features to maintain network connections. If a network connection to a node fails GPFS automatically tries to reestablish the connection before marking the node unavailable. This can provide for better uptime in environments communicating across a WAN or experiencing network issues.

Using these features along with a high availability infrastructure ensures a reliable enterprise class storage solution.

**GPFS Native Raid (GNR)**

Larger disk drives and larger file systems are creating challenges for traditional storage controllers. Current RAID 5 and RAID 6 based arrays do not address the challenges of Exabyte scale storage performance, reliability and management. To address these challenges GPFS Native RAID (GNR) brings storage device management into GPFS. With GNR GPFS can directly manage thousands of storage devices. These storage devices can be individual disk drives or any other block device eliminating the need for a storage controller.

GNR employs a de-clustered approach to RAID. The de-clustered architecture reduces the impact of drive failures by spreading data over all of the available storage devices improving application IO and recovery performance. GNR provides very high reliability through an 8+3 Reed Solomon based raid code that divides each block of a file into 8 parts and associated parity. This algorithm scales easily starting with as few as 11 storage devices and growing to over 500 per storage pod. Spreading the data over many devices helps provide predictable storage performance and fast recovery times measured in minutes rather than hours in the case of a device failure.
In addition to performance improvements GNR provides advanced checksum protection to ensure data integrity. Checksum information is stored on disk and verified all the way to the NSD client.

**Information lifecycle management (ILM) toolset**

GPFS can help you to achieve data lifecycle management efficiencies through policy-driven automation and tiered storage management. The use of storage pools, filesets and user-defined policies provide the ability to better match the cost of your storage to the value of your data.

Storage pools are used to manage groups of disks within a file system. Using storage pools you can create tiers of storage by grouping disks based on performance, locality or reliability characteristics. For example, one pool could contain high performance solid state disk (SSD) disks and another more economical 7,200 RPM disk storage. These types of storage pools are called internal storage pools. When data is placed in or moved between internal storage pools all of the data management is done by GPFS. In addition to internal storage pools GPFS supports external storage pools. External storage pools are used to interact with an external storage management application including IBM Tivoli Storage Manager (TSM) and High Performance Storage System (HPSS). When moving data to an external pool GPFS handles all of the metadata processing then hands the data to the external application for storage on alternate media, tape for example. When using TSM or HPSS data can be retrieved from the external storage pool on demand, as a result of an application opening a file or data can be retrieved in a batch operation using a command or GPFS policy. A fileset is a subtree of the file system namespace and provides a way to partition the namespace into smaller, more manageable units.
Filesets provide an administrative boundary that can be used to set quotas, take snapshots, define AFM relationships and be used in user defined policies to control initial data placement or data migration. Data within a single fileset can reside in one or more storage pools. Where the file data resides and how it is managed once it is created is based on a set of rules in a user defined policy. There are two types of user defined policies in GPFS: file placement and file management. File placement policies determine in which storage pool file data is initially placed. File placement rules are defined using attributes of a file known when a file is created such as file name, fileset or the user who is creating the file. For example a placement policy may be defined that states 'place all files with names that end in .mov onto the near-line SAS based storage pool and place all files created by the CEO onto the SSD based storage pool' or 'place all files in the fileset 'development' onto the SAS based storage pool'.

Once files exist in a file system, file management policies can be used for file migration, deletion, changing file replication status or generating reports.

You can use a migration policy to transparently move data from one storage pool to another without changing the file’s location in the directory structure. Similarly you can use a policy to change the replication status of a file or set of files, allowing fine grained control over the space used for data availability.

You can use migration and replication policies together, for example a policy that says: ‘migrate all of the files located in the subdirectory/database/payroll which end in *.dat and are greater than 1 MB in size to storage pool #2 and un-replicate these files’.

File deletion policies allow you to prune the file system, deleting files as defined by policy rules. Reporting on the contents of a file system can be done through list policies. List policies allow you to quickly scan the file system metadata and produce information listing selected attributes of candidate files.

File management policies can be based on more attributes of a file than placement policies because once a file exists there is more known about the file. For example file placement attributes can utilize attributes such as last access time, size of the file or a mix of user and file size. This may result in policies like: ‘Delete all files with a name ending in .temp that have not been accessed in the last 30 days’, or ‘Migrate all files owned by Sally that are larger than 4GB to the SATA storage pool’.

Rule processing can be further automated by including attributes related to a storage pool instead of a file using the threshold option. Using thresholds you can create a rule that moves files out of the high performance pool if it is more than 80% full, for example. The threshold option comes with the ability to set high, low and pre-migrate thresholds. Pre-migrated files are files that exist on disk and are migrated to tape. This method is typically used to allow disk access to the data while allowing disk space to be freed up quickly when a maximum space threshold is reached. This means that GPFS begins migrating data at the high threshold, until the low threshold is reached. If a pre-migrate threshold is set GPFS begins copying data until the pre-migrate threshold is reached. This allows the data to continue to be accessed in the original pool until it is quickly deleted to free up space the next time the high threshold is reached. Thresholds allow you to fully utilize your highest performance storage and automate the task of making room for new high priority content.

Policy rule syntax is based on the SQL 92 syntax standard and supports multiple complex statements in a single rule enabling powerful policies. Multiple levels of rules can be applied to a
file system, and rules are evaluated in order for each file when the policy engine executes allowing a high level of flexibility. GPFS provides unique functionality through standard interfaces, an example of this is extended attributes. Extended attributes are a standard POSIX facility. GPFS has long supported the use of extended attributes, though in the past they were not commonly used, in part because of performance concerns. In GPFS 3.4, a comprehensive redesign of the extended attributes support infrastructure was implemented, resulting in significant performance improvements. In GPFS 3.5, extended attributes are accessible by the GPFS policy engine allowing you to write rules that utilize your custom file attributes. Executing file management operations requires the ability to efficiently process the file metadata. GPFS includes a high performance metadata scan interface that allows you to efficiently process the metadata for billions of files. This makes the GPFS ILM toolset a very scalable tool for automating file management. This high performance metadata scan engine employs a scale-out approach. The identification of candidate files and data movement operations can be performed concurrently by one or more nodes in the cluster. GPFS can spread rule evaluation and data movement responsibilities over multiple nodes in the cluster providing a very scalable, high performance rule processing engine.

**Cluster configurations**

GPFS supports a variety of cluster configurations independent of which file system features you use. Cluster configuration options can be characterized into three basic categories:
- Shared disk
- Network block I/O
- Synchronously sharing data between clusters.
- Asynchronously sharing data between clusters.
**Shared disk**

A shared disk cluster is the most basic environment. In this configuration, the storage is directly attached to all machines in the cluster as shown in Figure 1. The direct connection means that each shared block device is available concurrently to all of the nodes in the GPFS cluster. Direct access means that the storage is accessible using a SCSI or other block level protocol using a SAN, InfiniBand, iSCSI, Virtual IO interface or other block level IO connection technology.

Figure 1 illustrates a GPFS cluster where all nodes are connected to a common fibre channel SAN. This example shows a fibre channel SAN though the storage attachment technology could be InfiniBand, SAS, FCoE or any other. The nodes are connected to the storage using the SAN and to each other using a LAN. Data used by applications running on the GPFS nodes flows over the SAN and GPFS control information flows among the GPFS instances in the cluster over the LAN.

This configuration is optimal when all nodes in the cluster need the highest performance access to the data. For example, this is a good configuration for providing network file service to client systems using clustered NFS, high-speed data access for digital media applications or a grid infrastructure for data analytics.

**Network-based block IO**

As data storage requirements increase and new storage and connection technologies are released a single SAN may not be a sufficient or appropriate choice of storage connection technology. In environments where every node in the cluster is not attached to a single SAN, GPFS makes use of an integrated network block device capability. GPFS provides a block level interface over TCP/IP networks called the Network Shared Disk (NSD) protocol. Whether using the NSD protocol or a direct attachment to the SAN the mounted file system looks the same to the application, GPFS transparently handles I/O requests.

GPFS clusters can use the NSD protocol to provide high speed data access to applications running on LAN-attached nodes. Data is served to these client nodes from one or more NSD servers. In this configuration, disks are attached only to the NSD servers. Each NSD server is attached to all or a portion of the disk collection. With GPFS you can define up to eight NSD servers per disk and it is recommended that at least two NSD servers are defined for each disk to avoid a single point of failure.

GPFS uses the NSD protocol over any TCP/IP capable network fabric. On Linux GPFS can use the VERBS RDMA protocol on compatible fabrics (such as InfiniBand) to transfer data to NSD clients. The network fabric does not need to be dedicated to GPFS; but should provide sufficient bandwidth to meet your GPFS performance expectations and for applications which share the
General Parallel File System (GPFS)

Technologies That Enable the Management of Big Data

bandwidth. GPFS has the ability to define a preferred network subnet topology, for example designate separate IP subnets for intra-cluster communication and the public network. This provides for a clearly defined separation of communication traffic and allows you to increase the throughput and possibly the number of nodes in a GPFS cluster. Allowing access to the same disk from multiple subnets means that all of the NSD clients do not have to be on a single physical network. For example you can place groups of clients onto separate subnets that access a common set of disks through different NSD servers so not all NSD servers need to serve all clients. This can reduce the networking hardware costs and simplify the topology reducing support costs, providing greater scalability and greater overall performance.

An example of the NSD server model is shown in Figure 2. In this configuration, a subset of the total node population is defined as NSD server nodes. The NSD Server is responsible for the abstraction of disk datablocks across a TCP/IP or Infiniband VERBS (Linux only) based network. The fact that the disks are remote is transparent to the application. Figure 2 shows an example of a configuration where a set of compute nodes are connected to a set of NSD servers using a high-speed interconnect or an IP-based network such as Ethernet. In this example, data to the NSD servers flows over the SAN and both data and control information to the clients flow across the LAN. Since the NSD servers are serving data blocks from one or more devices data access is similar to a SAN attached environment in that data flows from all servers simultaneously to each client. This parallel data access provides the best possible throughput to all clients. In addition it provides the ability to scale up the throughput even to a common data set or even a single file.
The choice of how many nodes to configure as NSD servers is based on performance requirements, the network architecture and the capabilities of the storage subsystems. High bandwidth LAN connections should be used for clusters requiring significant data transfer rates. This can include 1Gbit or 10 Gbit Ethernet. For additional performance or reliability you can use link aggregation (EtherChannel or bonding), networking technologies like source based routing or higher performance networks such as InfiniBand.

The choice between SAN attachment and network block I/O is a performance and economic one. In general, using a SAN provides the highest performance; but the cost and management complexity of SANs for large clusters is often prohibitive. In these cases network block I/O provides an option.

Network block I/O is well suited to grid computing and clusters with sufficient network bandwidth between the NSD servers and the clients. For example, an NSD protocol based grid is effective for web applications, supply chain management or modeling weather patterns.

**Mixed Clusters**

The last two sections discussed shared disk and network attached GPFS cluster topologies. You can mix these storage attachment methods within a GPFS cluster to better matching the IO requirements to the connection technology.

A GPFS node always tries to find the most efficient path to the storage. If a node detects a block device path to the data it is used. If there is no block device path then the network is used. This capability can be leveraged to provide additional availability. If a node is SAN attached to the storage and there is an HBA failure, for example, GPFS can fail over to using the network path to the disk. A mixed cluster topology can provide direct storage access to non-NSD server nodes for high performance operations including back-ups or data ingest.

**Sharing data between clusters**

There are two methods available to share data across GPFS clusters: GPFS multi-cluster and a new feature called Active File Management (AFM).

GPFS Multi-cluster allows you to utilize the native GPFS protocol to share data across clusters. Using this feature you can allow other clusters to access one or more of your file systems and you can...
General Parallel File System (GPFS)

Technologies That Enable the Management of Big Data

mount file systems that belong to other GPFS clusters for which you have been authorized. A multi-cluster environment allows the administrator to permit access to specific file systems from another GPFS cluster. This feature is intended to allow clusters to share data at higher performance levels than file sharing technologies like NFS or CIFS. It is not intended to replace such file sharing technologies which are optimized for desktop access or for access across unreliable network links. Multi-cluster capability is useful for sharing across multiple clusters within a physical location or across locations. Clusters are most often attached using a LAN, but in addition the cluster connection could include a SAN. Figure 3 illustrates a multi-cluster configuration with both LAN and mixed LAN and SAN connections. In Figure 3, Cluster B and Cluster C need to access the
data from Cluster A. Cluster A owns the storage and manages the
file system. It may grant access to file systems which it manages
to remote clusters such as Cluster B and Cluster C. In this example,
Cluster B and Cluster C do not have any storage but that is not a
requirement. They could own file systems which may be accessible
outside their cluster. Commonly in the case where a cluster does
not own storage, the nodes are grouped into clusters for ease of
management.

When the remote clusters need access to the data, they mount the
file system by contacting the owning cluster and passing required
security checks. In Figure 3, Cluster B accesses the data through the
NSD protocol. Cluster C accesses data through an extension of the
SAN. For cluster C access to the data is similar to the nodes in the
host cluster. Control and administrative traffic flows through the IP
network shown in Figure 3 and data access is direct over the SAN.
Both types of configurations are possible and as in Figure 3 can be
mixed as required.

Multi-cluster environments are well suited to sharing data across
clusters belonging to different organizations for collaborative
computing, grouping sets of clients for administrative purposes or
implementing a global namespace across separate locations.
A multi-cluster configuration allows you to connect GPFS clusters
within a data center, across campus or across reliable WAN links.
For sharing data between GPFS clusters across less reliable WAN
links or in cases where you want a copy of the data in multiple loca-
tions you can use a new feature introduced in GPFS 3.5 called Active
File Management.

Active File Management (AFM) allows you to create associations
between GPFS clusters. Now the location and flow of file data be-
tween GPFS clusters can be automated. Relationships between
GPFS clusters using AFM are defined at the fileset level. A fileset in
a file system can be created as a “cache” that provides a view to a
file system in another GPFS cluster called the “home.” File data is
moved into a cache fileset on demand. When a file is read the in the
cache fileset the file data is copied from the home into the cache fi-
leset. Data consistency and file movement into and out of the cache
is managed automatically by GPFS.

Cache filesets can be read-only or writeable. Cached data is locally
read or written. On read if the data is not in the “cache” then
GPFS automatically creates a copy of the data. When data is writ-
ten into the cache the write operation completes locally then GPFS
asynchronously pushes the changes back to the home location.

You can define multiple cache filesets for each home data source.
The number of cache relationships for each home is limited only
by the bandwidth available at the home location. Placing a quota
on a cache fileset causes the data to be cleaned (evicted) out of
the cache automatically based on the space available. If you do not set
a quota a copy of the file data remains in the cache until manually
evicted or deleted.
You can use AFM to create a global namespace within a data center, across a campus or between data centers located around the world. AFM is designed to enable efficient data transfers over wide area network (WAN) connections. When file data is read from home into cache that transfer can happen in parallel within a node called a gateway or across multiple gateway nodes. Using AFM you can now create a truly world-wide global namespace. Figure 5 is an example of a global namespace built using AFM. In this example each site owns 1/3 of the namespace, store3 “owns” /data5 and /data6 for example. The other sites have cache filesets that point to the home location for the data owned by the other clusters. In this example store3 /data1-4 cache data owned by the other two clusters. This provides the same namespace within each GPFS cluster providing applications one path to a file across the entire organization. You can achieve this type of global namespace with multi-cluster or AFM, which method you chose depends on the quality of the WAN link and the application requirements.
Figure 5: Global Namespace using AFM

Client access:
- /global/data1
- /global/data2
- /global/data3
- /global/data4
- /global/data5
- /global/data6

Cache: /data1
Local: /data3
Cache: /data5

Store 1

Client access:
- /global/data1
- /global/data2
- /global/data3
- /global/data4
- /global/data5
- /global/data6

Cache: /data1
Local: /data3
Cache: /data5

Store 2

Client access:
- /global/data1
- /global/data2
- /global/data3
- /global/data4
- /global/data5
- /global/data6

Cache: /data3

Store 3

Client access:
- /global/data1
- /global/data2
- /global/data3
- /global/data4
- /global/data5
- /global/data6

Cache: /data1
Local: /data3
General Parallel File System (GPFS)

What’s new in GPFS Version 3.5

What’s new in GPFS Version 3.5
For those who are familiar with GPFS 3.4 this section provides a list of what is new in GPFS Version 3.5.

Active File Management
When GPFS was introduced in 1998 it represented a revolution in file storage. For the first time a group of servers could share high performance access to a common set of data over a SAN or network. The ability to share high performance access to file data across nodes was the introduction of the global namespace. Later GPFS introduced the ability to share data across multiple GPFS clusters. This multi-cluster capability enabled data sharing between clusters allowing for better access to file data. This further expanded the reach of the global namespace from within a cluster to across clusters spanning a data center or a country. There were still challenges to building a multi-cluster global namespace. The big challenge is working with unreliable and high latency network connections between the servers. Active File Management (AFM) in GPFS addresses the WAN bandwidth issues and enables GPFS to create a world-wide global namespace. AFM ties the global namespace together asynchronously providing local read and write performance with automated namespace management. It allows you to create associations between GPFS clusters and define the location and flow of file data.

High Performance Extended Attributes
GPFS has long supported the use of extended attributes, though
in the past they were not commonly used, in part because of performance concerns. In GPFS 3.4, a comprehensive redesign of the extended attributes support infrastructure was implemented, resulting in significant performance improvements. In GPFS 3.5, extended attributes are accessible by the GPFS policy engine allowing you to write rules that utilize your custom file attributes. Now an application can use standard POSIX interfaces to manage extended attributes and the GPFS policy engine can utilize these attributes.

### Independent Filesets
To effectively manage a file system with billions of files requires advanced file management technologies. GPFS 3.5 introduces a new concept called the independent fileset. An independent fileset has its own inode space. This means that an independent fileset can be managed similar to a separate file system but still allow you to realize the benefits of storage consolidation.

An example of an efficiency introduced with independent filesets is improved policy execution performance. If you use an independent fileset as a predicate in your policy, GPFS only needs to scan the inode space represented by that fileset, so if you have 1 billion files in your file system and a fileset has an inode space of 1 million files, the scan only has to look at 1 million inodes. This instantly makes the policy scan much more efficient.

Independent filesets enable other new fileset features in GPFS 3.5.

### Fileset Level Quotas
User and group quotas can be set per fileset.

### File Cloning
File clones are space efficient copies of a file where two instances of a file share data they have in common and only changed blocks require additional storage. File cloning is an efficient way to create a copy of a file, without the overhead of copying all of the data blocks.

### I Pv6 Support
IPv6 support in GPFS means that nodes can be defined using multiple addresses, both IPv4 and IPv6.

### GPFS Native RAID
GPFS Native RAID (GNR) brings storage RAID management into the GPFS NSD server. With GNR GPFS directly manages JBOD based storage. This feature provides greater availability, flexibility and performance for a variety of application workloads.

GNR implements a Reed-Solomon based de-clustered RAID technology that can provide high availability and keep drive failures from impacting performance by spreading the recovery tasks over all of the disks. Unlike standard Network Shared Disk (NSD) data access GNR is tightly integrated with the storage hardware. For GPFS 3.5 GNR is available on the IBM Power 775 Supercomputer platform.
General Parallel File System (GPFS)

GPFS Storage Server – a home for Big Data

- IBM System x storage solution building block with
- 2 servers with Linux and IBM GPFS – real POSIX file system
- today 4 or 6 disk enclosures with 232x/348x 2 or 3 TB NL-SAS Enterprise disks.
- Standard 42U 19 inch Rack
- 10 Gigabit Ethernet or FDR Infiniband interfaces
- Physical installation included
- 3 years warranty included
- Software implementation as service offering

Value Proposition

- **Reliability**: most reliable file system in all redundant packaging; frequent failure source “Controller” eliminated by declustered Software-RAID
- **Integrity**: only integrated solution with End-to-End Checksums including versioning for the safest identification and correction of silent data corruption/dropped writes
- **Performance**: with 10GB/s and 12GB/s per building block highest performance density, near linear and unlimited (>1 YobiByte) scalable in one file system and name space
**Big Data** – hundreds of TeraByte and PetaByte of live data

- Media data in streaming, game development, movie editing
- Connection data for transport, carriers, ...
- Connection data, hosting, cloud storage for internet and communication companies
- Data of digitalization projects, administration data
- Patent data, Tax data and more on the public sector
- User/Home data in Higher Education

**Technical Computing** – Live data in analytics and simulation

- Health care and biology data like Genomes, Scans, ...
- High Performance Computing “Scratch”-storage
- Data for risk analysis, Basel III CVA data and more
- Model data in CAE
ALWAYS KEEP IN TOUCH WITH THE LATEST NEWS

Visit us on the Web!

Here you will find comprehensive information about HPC, IT solutions for the datacenter, services and high-performance, efficient IT systems.

Subscribe to our technology journals, E-News or the transtec newsletter and always stay up to date.

www.transtec.de/en/hpc