



ALPS

Advanced Large Parallel System

ALPS Supercomputing System
A Scalable Supercomputer with Flexible Services



Abstract

Supercomputing is moving from the realm of abstract to mainstream with more and more applications and research being ported to the digital environment. These applications can range from manufacturing, to oil and gas exploration, to molecular biology, to simulating weather and natural disasters, and even 3D rendering for feature films. These high-performance computing (HPC) systems are being used to tackle increasing larger amounts of workloads, and are more necessary to answer the problems of today than ever before.

Taiwan's National Center for High-performance Computing (NCHC) is the leading research institute on the island for all things HPC. The center works with academic and enterprise institutes to take part in joint research and HPC service support in the fields of biotechnology, energy resources, water resources, life sciences and medicine, chemistry and more. To further their service and support for the growing HPC research needs, the center sought to build a new supercomputing system that can offer scalable HPC resources to many institutions across the island.

Code-named ALPS, the new system is built by Acer Group and its key HPC partners, including Qlogic, DataDirect Networks, Platform Computing and Alinea, and offers an aggregate performance of over 177 TFLOPS. The system uses the latest AMD Opteron™ processors, and has a total of 8 compute clusters, 1 large memory cluster, and over 25,000 compute cores.

This white paper details many of the new features and challenges which were overcome in the development of this all-new system, and outlines the applications NCHC seeks to address.

Building a scalable supercomputer

Contents

System overview

- Specifications
- Topology
- Storage breakdown

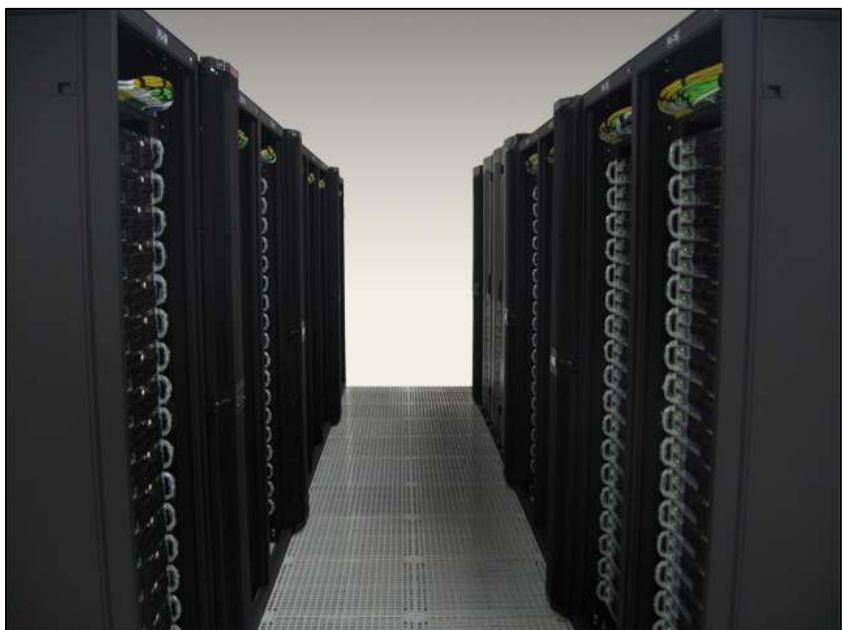
Multi-cores become 'Many-core' – AMD Opteron™ 6100 processors

High-performance InfiniBand Networking

- Qlogic TrueScale solution

Advanced cluster and workload management

- Platform HPC management and performance benefits



Scalable Supercomputing – system overview

Supercomputing has evolved quite a bit from the massive mainframe and supercomputing systems of just a little over 10 years before. Since cluster computing became mainstream, most supercomputing systems developed now are based upon the concept of parallel computing. In a parallel setup information and 'jobs' are given to a login node (head node), broken down into pieces, and then distributed to a group of computers (compute nodes) for processing. The advantage of such a design is that jobs which would have once been impossible to compute in a reasonable amount of time on a single computer, can quickly be broken down and processed. Additionally, because the systems are based on smaller pieces of hardware there exists a much greater degree of flexibility in terms of what can be customized and scaled.

When creating a flexible system it is important to take into account several factors, not the least of which is current and future compatibility. Simply put, if the system is not compatible with the software or other applications that needs to be run, no matter the performance, the system is worthless.

NCHC provides supercomputing services to a variety of academic institutions in Taiwan. Its flagship system, ALPS Supercomputing System *Windrider*, had to be designed to support a wide range of applications – so many that NCHC could only begin to fathom the jobs that could potentially be run across it. One day a part of the cluster could be dedicated to molecular chemistry research, while another day could be financial calculations, or even seismic research of the surrounding region.

The new system is also designed to function as a test bed for new application design and research; support for open source and academic-specific applications is a must. To offer the most robust level of support, the system was first planned with every angle in mind – not simply raw compute performance, and not necessarily best performance per watt.

To meet these demands, Acer and its partners proposed a robust platform that offered maximum performance in terms of processing power, storage I/O throughput, memory size and general compatibility. Though planning for a narrower range of applications could open up the possibility for other solutions, such as GPU or Unix, the chosen design consists 100% of the latest AMD Opteron™ 6100 processors that support older and newer applications equally well.

Storage-wise, it is generally better to provide too much I/O throughput than too little, and since the system needs to support an unknown range of applications, reaching a balance between I/O and CPU performance was a must. Acer Group's HPC storage partner, DataDirect Networks (DDN) provides first class storage performance and support via its own Lustre-based parallel file system. Already powering several HPC systems around the globe, DDN was the natural choice for Acer when planning the hardware layout for the build. The result was a storage system that is one of the largest in terms of throughput in Asia – up to 9.3 GB/s dedicated write and read performance for any sub-cluster of 64 nodes.

The cluster interconnect network is a vital component of the entire system with low message latency and high bandwidth being necessary to allow applications to scale in performance as the number of compute nodes increases. Acer chose to work with the networking experts at QLogic to design a powerful high-speed network based on QLogic's TrueScale InfiniBand products, capable of keeping up with the MPI and I/O traffic needs of the system. The resulting system is a dual-switch, independent MPI and I/O network design allowing for ample amounts of network bandwidth – over 100 Tb/s aggregate.

Platform Computing brings a complete cluster and workload management solution to the system with its Platform HPC product. Platform HPC incorporates a powerful job scheduler based on Platform LSF, a robust MPI application library, Platform MPI, as well as comprehensive cluster management features, advanced monitoring and reporting, and easy-to-use web interface for job submission, Platform HPC provides a unified set of management capabilities that allows you to deploy, manage and maintain the

Building a scalable supercomputer

clustered hardware, as well as ensuring maximum throughput and utilization of the cluster to provide optimal return for the investment made by NCHC.

Specifications overview

ALPS highlights:

- 25,000+ computing cores
- 73 +TB DDR3 memory
- 1000+ TB usable storage
- **177 TFLOPS of performance**



Structure	8 x compute clusters with 64 compute nodes and 3 available head nodes per cluster 1 x large memory compute cluster with 32 compute nodes and 2 head nodes
Compute	536 x AR585 F1 with 4 x AMD Opteron 6174 (12 core, 2.2 GHz) 34 x AR585 F1 with 4 x AMD Opteron 6136 (8 core, 2.4 GHz)
Interconnect	Qlogic TrueScale (QDR) InfiniBand with 40 Gb/s of bandwidth via a single port 2 x Qlogic 12800-360 director IB switches with up to 640 fully non-blocking ports and 51.8 Tb/s of bandwidth per switch
Parallel file system	10 x DataDirect Networks (DDN) Exascaler Lustre file systems 1- x DDN SFA10000 storage systems with dedicated SSD metadata storage
Cluster management	Platform Computing – Platform HPC 2.1 providing: <ul style="list-style-type: none"> • Cluster management • Workload management • Workload monitoring and reporting • System monitoring and reporting • Dynamic OS multi-boot • Commercially supported MPI libraries • Integrated application portal with prebuilt templates and scripts • Unified web management portal
Compiler	PGI Server Compiler Intel Cluster Studio
Operating system	Novell SUSE 11 SP1 Enterprise Server

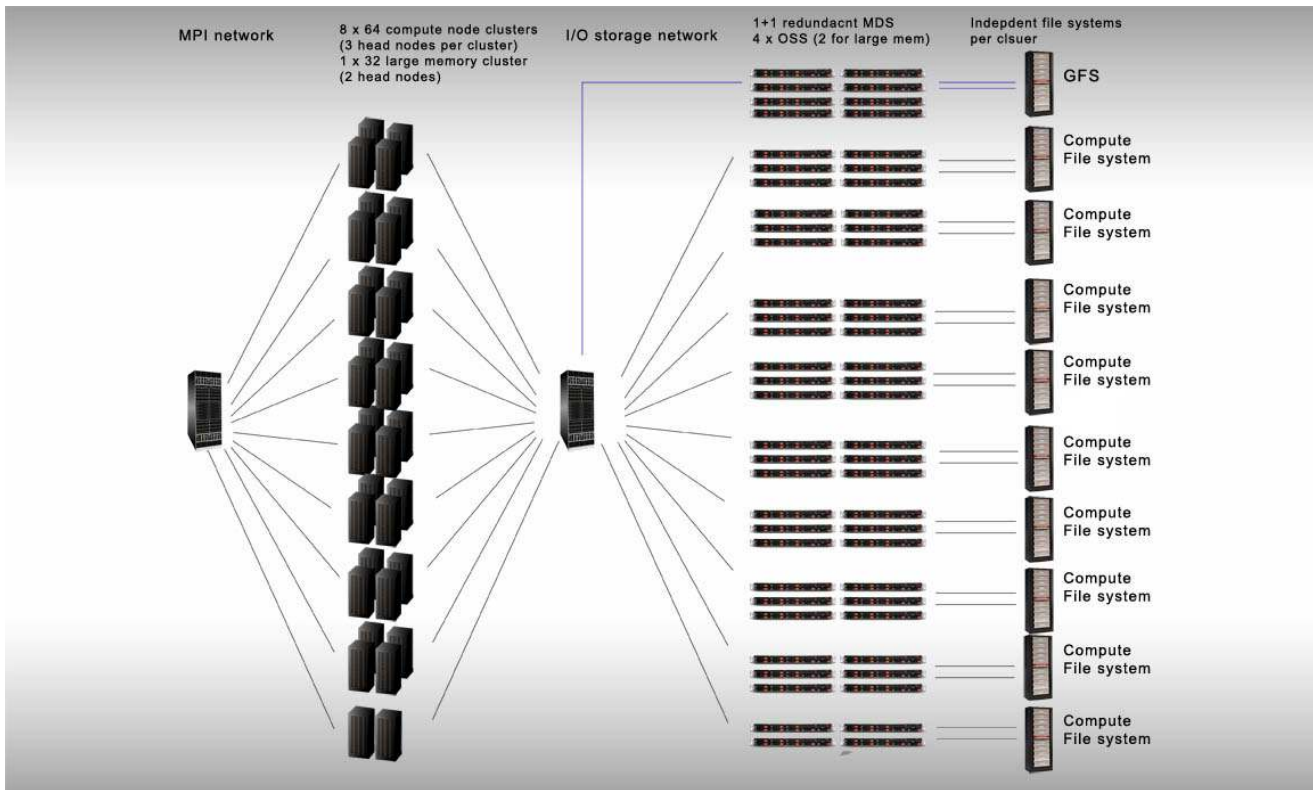


Building a scalable supercomputer

Topology

ALPS is a scalable system meaning it can be expanded or shrunk to fit any research need. As such the topology is flexible.

The original setup includes 8 sub clusters of computing nodes and 1 smaller cluster for large memory. Platform Computing's Platform HPC's cluster and workload management capabilities make it easy for users to harness the power and scalability of their HPC cluster resulting in shorter time for system readiness and user productivity as well as optimal throughput.



The high-speed interconnect is completely split between I/O and MPI traffic to ensure maximum performance regardless of the job. Both types of traffic use QDR InfiniBand with up to 40 Gb/s of bandwidth, and this interconnect goes directly to the storage controller.

Simultaneously, by having the topology grouped into two large switches, the configuration becomes even more configurable and customizable. Changing the compute nodes into separate clusters, and even combining clusters can all be handled within the Platform environment. Compute nodes can be moved from one cluster to another in under five commands, and no re-wiring is required. This flexibility allows NCHC to increase, combine or decrease any cluster capacity whenever they want which is ideal for an HPC solution.

Storage topology – parallel file system

Different from local disks, block storage arrays provide RAID data protection and high availability for mission-critical systems. However, in the case of large compute cluster, this is not enough; clustered systems are best run as a single system sharing everything from cores to memory to storage. In order to achieve this unity with the best performance, the storage must run in parallel across multiple connections. A

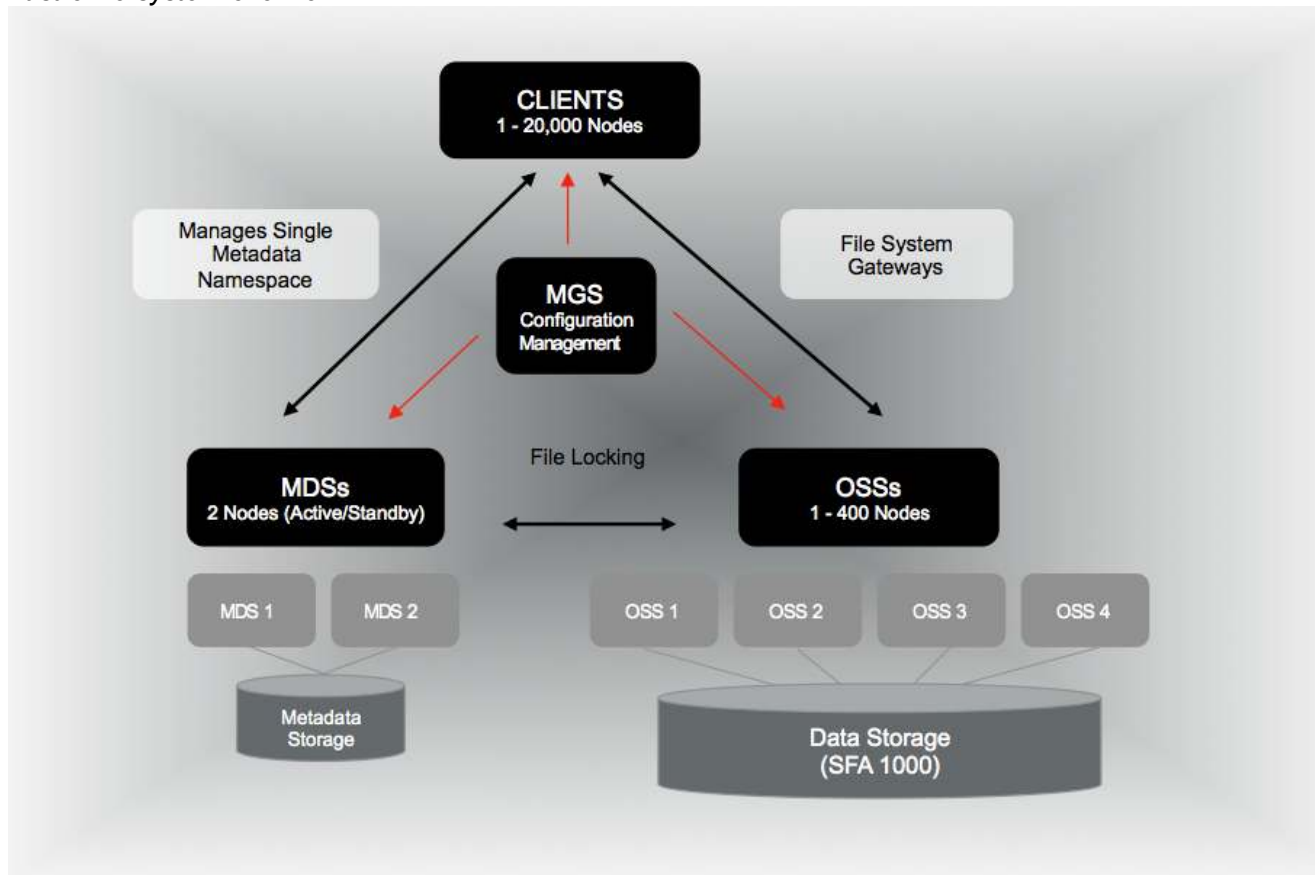
Building a scalable supercomputer

scalable parallel file system is required to achieve performance levels well beyond those than can be achieved by local storage or alternatives, such as the Network File System (NFS).

The advantage is obvious: all nodes have equal access to data — significantly simplifying the handling of data needed for computation. When all nodes see the same storage environment, complex data staging is no longer necessary. Further, parallel access to the same file system increases performance for most HPC environments.

The open source Lustre file system offers the scalability needed for large-scale clustered supercomputers. In a Lustre environment, output can scale by increasing the balance of object storage servers (OSS) and backend storage systems with highly optimized disk arrays. When paired with QDR InfiniBand, a single OSS is able to produce up to around 2.8 GB/s of throughput to the cluster, and this can scale up as more OSS are added. Backend performance and enterprise class data protection is provided by DDN's Storage Fusion Architecture (SFA) 10000 storage arrays.

Lustre file system overview



Source: ExaScaler "LUSTRE™ FILE SYSTEM, DDN 2011.

The ALPS supercomputing system employs ten Lustre file systems across the nine clusters – a single file system for each cluster and a global file system accessible to all clusters on top. Thus, the clusters can act as 100% independent units if needed or may be grouped together into larger clusters for more intensive applications.

DataDirect Networks is a worldwide leader in scalable, performance and capacity driven storage systems. ExaScaler, the technical storage solutions integrated by DDN includes a pre-packaged version of the

Building a scalable supercomputer

Lustre file system that has undergone stress testing at numerous HPC sites and provides remote monitoring, setup and re-configuration services to the ALPS cluster. Using the SFA10000 storage appliance combined with DDN's Lustre-based ExaScaler solution, DDN achieved over 3 MB/s/core of throughput for any given cluster of over 3000 compute cores. In total, the file system environment for ALPS is capable of writing or reading over 100 GB of data per second.

DDN provides support for the ExaScaler file system in partnership with Whamcloud, the company that is the leading developer of the Lustre file system.

Multi-cores become 'many-cores'

In the past five years, multi-core processors have become mainstream, being used in everything from servers to PCs and even smart phones. The advantages are numerous, offering more processing power without the need to push for greater frequency levels – a technique that often requires more power and offers only limited performance increase.

In 2010, AMD released its latest server-level CPU platform, AMD Opteron™ 6100 series, that broke the previous multi-core barrier and brought enterprise computing into the realm of 'many-cores'. With a maximum of 12 cores per CPU, the latest AMD Opteron 6100 series of processors is a perfect fit for a parallel computing environment designed to leverage as many individual cores as possible. High-performance computing is just such an environment, and is able to fully leverage incredibly high core counts.

AMD's Opteron 6100 series is also a departure from previous server platform designs as it can scale from two socket to four socket server boards without having to migrate to premium-priced processors. In so doing, a homogenous four-socket architecture, like that used in ALPS, was even more scalable in terms of both price and performance.

Leveraging AMD's four socket design with up to 48 cores in a single platform, Acer created a 25,000 core system that was able to scale with an efficiency of over 75 percent and decrease power and chassis costs dramatically. The high-core count system is not only ideal for parallel programming, but provides NCHC with even more flexibility to break down their system into a greater number of sub-clusters and handle dozens or even hundreds of jobs in a truly parallel fashion.

Key features of the AMD Opteron 6100 series and platform



- Up to twelve cores per processor, perfect for handling scalable HPC applications like computation fluid dynamics, life sciences, computational chemistry, weather modeling and more, double the competition
- Four DDR-3 memory channels deliver up to 54GB/s of memory throughput in STREAM benchmarks, more than 45% more memory throughput than the competition
- Up to 12 128-bit floating point units per CPU, giving an average performance of over 340 GFLOPs per AR585 F1 in the ALPS Supercomputing System

The flexibility of the new AMD Opteron 6100 Processor is not limited to performance. The power efficiency of this architecture can offer a really extreme dense platform with up to 48 CPU cores in a 1U form factor for the best server footprint, and the possibility to configure up to 8 DIMMs per CPU (on the AR585 used in the project) on four independent memory channels provide an incredible amount of memory availability for HPC applications, extremely important when working with large data sets like weather modeling and computational fluid dynamics.

Building a scalable supercomputer

AMD processors power 24 of the world's top 50 fastest supercomputers ranked in the bi-annual TOP500[®] list, and its latest enterprise class 2-way and 4-way AMD Opteron™ 6000™ Series server platforms provide consistent, energy efficient server platforms that scale to any HPC solution.¹

High-performance InfiniBand networking

In the past several years as proprietary interconnect networks have declined, InfiniBand[®] has emerged as the leading choice for major HPC interconnect fabrics providing the highest available bandwidth and lowest latency of the open/standards based interconnect technologies.

The QLogic TrueScale™ InfiniBand product family is a comprehensive end-to-end portfolio that delivers unsurpassed performance, scalability and cost effectiveness. The TrueScale portfolio includes:

- InfiniBand host channel adapters
- InfiniBand switches
- Advanced management software

The QLogic TrueScale host channel adapters which connect the cluster nodes to the InfiniBand network have been designed specifically for HPC applications. This may seem obvious but when InfiniBand was first specified the target market was the Enterprise Data Center which has different connectivity requirements to message passing applications found on HPC clusters. The adapters use a lightweight connectionless communications protocol and the processing is balanced between the host processor and adapter itself. This provides a very efficient interface between message passing application and the cluster interconnect which allows the communications performance to scale with the performance of the compute node and the number of nodes involved in the calculation. The benefit of QLogic's low end-to-end latency is also seen in an optimized MPI collectives performance.

The QLogic TrueScale 1200 series of switches, available in edge, rack or director class form factors, offer the highest port count and highest port density in the industry, as well as superior configuration topologies.

QLogic TrueScale also supports a number of advanced fabric services. Adaptive routing implemented in the switch hardware automatically manages the routing of messages around congested areas of the fabric resulting in a more efficient use of the fabric and higher communications performance of an individual application. Dispersive routing optimizes communications by splitting up messages and routing the parts across multiple routes before re-assembly at the destination node. Virtual fabrics enable cluster administrators to dedicate virtual lanes and classes of service by application within the network allowing predictable sharing of the network.

Advanced cluster and workload management

To ensure NCHC has the right software to run such a large and complex cluster, the ALPS system runs Platform HPC, the leading HPC Software solution for managing HPC environments. Platform HPC is a complete HPC management product, which ensures rapid system deployment, advanced infrastructure and applications management, and significant ease of use features for both Administrators and HPC Users.

Platform HPC incorporates powerful cluster management features, including multiple provisioning methods such as image based, package based and diskless, depending on customer requirements. NCHC has chosen to use a diskless configuration to increase overall performance, but could choose to use other methods of provisioning at any point in the future. Platform HPC manages all cluster node software, ensuring consistency across the entire system.

¹ Number based on the November 2010 report of the Top500. www.top500.org

Building a scalable supercomputer

Platform HPC also features Platform LSF, the world's leading HPC Workload Management product that ensures maximum throughput and utilization of all HPC resources. With its advanced scheduling policies that allow resources to be allocated and prioritized to those users and groups that take precedent over other users, Platform HPC ensures that deadlines and SLA are met.

Platform HPC also incorporates a web based job submission and management portal. This provides users with web based access to published applications, with jobs submitted using templates. These templates are completely customizable, allowing HPC Administrators to modify the way in which job details are submitted using drop down boxes and text fields. Users can also indicate the input data, and output locations to send their files on completion. This massively reduces the time to learn how to submit jobs to the cluster, ensuring that the system at NCHC is used as much as possible.

Platform HPC also includes Platform MPI, the leading commercial MPI library for parallel workloads. Platform tuned the MPI using the NCHC System, and these improvements were part of the reason the system delivered the 177 TFLOPS benchmark achieved.

To give NCHC the visibility of how the system is operating, and how it is being used, sophisticated monitoring and reporting keeps a watchful eye on all of the hardware and infrastructure. This includes the Acer servers, InfiniBand and Ethernet switches, and collates alerts and alarms into a single dashboard across the entire system. From there, administrators can drill down into individual sub clusters and then individual node. The reporting also tracks and reports on all workloads across the cluster, allowing NCHC to report on throughput and utilization by user, by application, by cluster, by group, and by many other metrics, ensuring managers are informed of how the resources are being utilized, and how efficiently the system is running.

ALPS Supercomputing System

The ALPS Supercomputing System is a powerful and flexible system that will meet the research needs across Taiwan and Asia region. The system will be open for public use beginning in mid-Q3 of 2011.

Copyright © 2011. All rights reserved.