

Appro **HyperPower**[™] Blade Cluster

- Hybrid, modular, and scalable solution
- Solution based on the Appro GreenBlade™ System
- Integrates the latest CPU and GPU technologies
- Delivers 8x faster performance than previous generation
- Flexible configurations that include management options
- Ideal for Parallel Data Processing in HPC applications



Appro HyperPower™ Blade Cluster

The amount of raw data needed for computational analysis and 3D visualization has created a huge demand for CPU/GPU cluster deployments.

With the need for more performance and memory capacity, Appro introduces the Appro HyperPower Blade Cluster offering a hybrid modular architecture based on the Appro GreenBlade System. This building block solution consolidates server, storage, network, power and simplified management capabilities while providing performance, reliability, density, upgradeability and value. Configurations are complemented with GPU expansion blades offering customers affordable CPU/GPU computing blade options in a single solution building block. The system comes in a 5U form factor and supports up to 10 server blades in one package. If configured as a hybrid CPU and GPU system, the Appro GreenBlade system supports 5 dual CPU server blades with 5 GPU expansion blades delivering up to 5 TFlops of double precision performance per system.

The HyperPower Blade cluster offers performance scalability, density and energy efficiency at a competitive price. It allows scientific and technical professionals the ability to develop applications faster and to deploy them across multiple generations of processors.

PetaScale Computing with TeraFlop Processors

The Appro HyperPower Blade Clusters based on NVIDIA® Tesla™ 20 series GPU computing enables the transition to energy efficient parallel computing power making Petascale computing with TeraFlop processors possible. Each compute server includes two Tesla GPUs, each offering 448 cores delivering 8x increase in double precision performance compared to Tesla 10-series. The Tesla™ 20 series GPUs are designed to redefine high performance computing and make supercomputing available to everyone.

NVIDIA Tesla 20 series delivers supercomputing performance at 1/10th the cost and 1/20th the power consumption. In addition, NVIDIA Tesla scales to solve the world's most important computing challenges-more quickly and accurately.

GPU Architecture Delivers Optimum Scaling Across HPC Applications

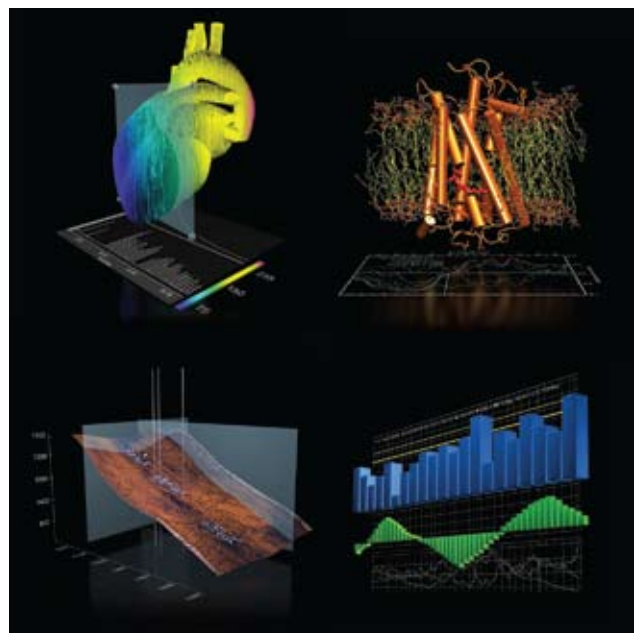
Scientists and engineers have made the transition to 'many-core' computing because their problems have reached a complexity that required them to look for new ways to boost their work. The revolutionary scalar, single, and double precision floating-point performance enables solving a wide range of high- performance computing applications, whose complexity has outstripped the CPU's ability to solve them.

NVIDIA® CUDA™ Technology

The NVIDIA CUDA architecture, codenamed "Fermi", enables parallel computing architecture to simplify many-core programming and enhances performance by offloading computationally-intensive activities from the CPU to the GPU. It enables developers to utilize NVIDIA GPUs to solve the most complex computation-intensive challenges such as protein docking, molecular dynamics, financial analysis, fluid dynamics, structural analysis and many others.

Ideal Environment

Ideal for small, mid and large-sized HPC deployments in Government, Research Labs, Universities and vertical industries such as Oil and Gas and Bioinformatics where the most computationally-intensive applications are needed.



Small to Mid-Size Data Centers Scaling Up To 1000 Nodes



Appro HyperPower™ Blade Cluster



Appro GreenBlade™ System



CPU Blade Expansion



Performance Optimized GPU Cluster

Appro HyperPower™ Blade Cluster offers a hybrid solution based on two Intel® Xeon® CPUs and two NVIDIA® Tesla™ M2050 GPU modules per blade node.

- Delivers equivalent supercomputing performance at 1/10th the cost and 1/20th the power consumption compared to CPU only clusters.
- Features a total of 80 blades in a 42U standard rack, 40x blades based on Intel® Xeon® processor 5600 series combined with 40x GPU expansion blades based on NVIDIA® Tesla™ M2050 GPU modules.
- Supports 480 CPU/35,840 GPU cores delivering up to 41TF of GPU double precision performance per rack.
- Supports a variety of configurations, interconnects and cluster management options with a choice of Linux or Windows® operating systems.
- Tested, pre-integrated solution deployed as a complete package.
- HPC professional service and support available.

Modular, Flexible and Reliable Design

The GreenBlade System offers customers affordable mix and match CPU/GPU compute blade options that are very easy to configure and deploy:

- Consolidates server, storage, network - all in a shared environment utilizing 90%+ high efficiency power supplies.
- Offers hot-swappable and redundant core components such as cooling fans, power supplies and blade nodes providing superior reliability, availability and serviceability.
- Supports up to 5 dual CPU server blades with 5 dual GPU expansion blades offering up to 4,480 GPU cores per system.
- Features Intelligent Power Control to dynamically power down pairs of GPUs to provide power savings while offering thermal efficiency and reliability.
- Integrated IPMI 2.0 remote server management capabilities.

CPU Performance That Adapts to Your Environment

Application performance is critical for day-to-day business operations, as well as creating new products and reaching new customers. But many data centers are now at capacity, and new ones are expensive to build. By refreshing data center infrastructure with more efficient servers, customers can deliver additional performance and scalability within the same energy and space footprint.

Appro HyperPower based on Intel® Xeon® processor 5600 series delivers intelligent performance:

- **Intel Intelligent Power Technology** makes power available for critical workloads while conserving power where there is less demand, delivering as much as 40 percent better performance in a similar power envelope.
- **Intel Turbo Boost Technology** increases performance by automatically increasing core frequencies and enabling faster speeds for specific threads and mega-tasking workloads.
- **Intel Hyper-Threading Technology** benefits from larger caches and massive memory bandwidth, delivering greater throughput and responsiveness for multi-threaded applications.
- **Intel QuickPath Technology** and integrated memory controller speed traffic between processors and I/O controllers for bandwidth-intensive applications, delivering up to 4.4x the bandwidth for technical computing.
- **Large Memory Capacity**
Up to 12 DIMM slots with up to 96GB of main memory for higher performance for your data intensive applications.
- **12MB Shared L3 Cache with Enhanced Smart Cache**
Inclusive shared L3 cache increases application performance by reducing traffic to the processor cores.
- **Intel® I/O Acceleration Technology**
Moves data more efficiently through Intel® Xeon® processor-based workstations and server platforms for fast, scalable, and reliable network performance.

World's Most Adaptable Server Platform

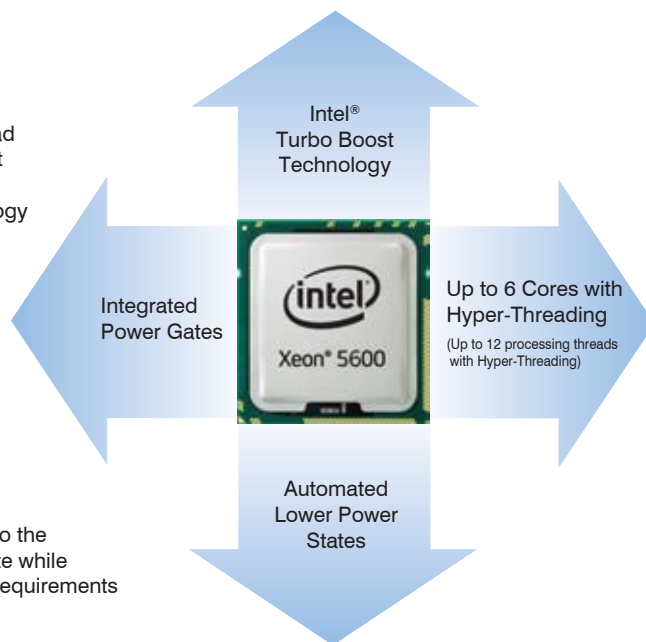
Intel® Xeon® Processor 5600 Series

Performance

Maximizes performance by adapting to the workload through Intel® Turbo Boost Technology and Intel® Hyper-Threading Technology

Software Adaptable

The processor adapts to the way your application wants to run



Energy Efficiency

Automatically puts CPU into the lowest available power state while still meeting performance requirements

IT Adaptable

You can enable automatic operation or selectively configure for manual control

GPU Feeding the Relentless Demand for HPC Performance

The Appro HyperPower Blade Cluster closes the gap between the demands placed by application performance and performance delivered by the computing processor. With the massively parallel architecture of the GPU, scientists and engineers can get a quantum leap in performance and continue to advance the pace of their work, guiding them to faster discovery in drug research, weather modeling, oil and gas exploration, computational finance, and more.

Tesla M2050/M2070 GPU Computing Module

448 CUDA Cores	Delivers up to 515 Gigaflops of double-precision peak performance in each GPU, enabling servers from Appro to deliver a Teraflop or more of double precision performance per 1 RU of space. Single precision peak performance is over one Teraflop per GPU.
ECC Memory	Meets a critical requirement for computing accuracy and reliability in datacenters and supercomputing centers. Offers protection of data in memory to enhance data integrity and reliability for applications. Register files, L1/L2 caches, shared memory, and DRAM all are ECC protected.
Up to 6GB of GDDR5 memory per GPU	Maximizes performance and reduces data transfers by keeping larger data sets in local memory that is attached directly to the GPU.
System Monitoring Features	Integrates the GPU subsystem with the host system's monitoring and management capabilities. This means IT staff can manage all of the critical components of the computing system through a common management interface such as IPMI or OEM-proprietary tools.
Designed for Maximum Reliability	Passive heatsink design eliminates moving parts and cables.
NVIDIA Parallel Data Cache™	Accelerates algorithms such as physics solvers, ray-tracing, and sparse matrix multiplication where data addresses are not known beforehand. This includes a configurable L1 cache per Streaming Multiprocessor block and a unified L2 cache for all of the processor cores.
NVIDIA Giga Thread™ Engine	Maximizes the throughput by faster context switching that is 10X faster than previous architecture, concurrent kernel execution, and improved thread block scheduling.
Asynchronous Transfer	Turbocharges system performance by transferring data over the PCIe bus while the computing cores are crunching other data. Even applications with heavy data-transfer requirements, such as seismic processing, can maximize the computing efficiency by transferring data to local memory before it is needed.
CUDA programming environment with broad support of programming languages and APIs	Choose C, C++, OpenCL, DirectCompute, or Fortran to express application parallelism and take advantage of the innovative "Fermi" architecture.
High Speed, PCIe Gen 2.0 Data Transfer	Maximizes bandwidth between the host system and the Tesla processors. Enables Tesla systems to work with virtually any PCIe-compliant host system

Supported Operating Systems

- Linux® 64-bit and 32-bit
- Red Hat Enterprise Linux 5
- SUSE 10.2 and 10.3
- Windows® Server 2003 and 2008

Software Development Tools

CUDA C/C++/Fortran, OpenCL, DirectCompute Toolkits.
 NVIDIA Parallel Nsight™ for Visual Studio



HyperPower™ Blade Cluster Configuration

HyperPower Blade Cluster

Blades	Up to 80 blades per rack - 40x CPU and 40x GPU
Processors	Intel® Xeon® 5500/5600 series CPU and NVIDIA® Tesla® M2050 GPUs
CPU/GPU Cores	Up to 480 CPU/35,840 GPU cores
Memory Capacity	Up to 3.84TB CPU ECC DDR3 memory Up to 240GB GPU GDDR5 memory
Performance	46.83TF double-precision (CPU & GPU total performance) Up to 5.63TF CPU performance (with X5670 CPUs) Up to 82.4TF GPU single-precision performance Up to 41.2TF GPU double-precision performance
Storage Capacity	Up to 80 internal 2.5" HDDs, equal to 40TB of local storage
Networking	Standard: Gigabit Ethernet Optional: Integrated QDR InfiniBand™
Rack Configuration	Standard 42U/19" rack, 2U rack space available for switches
RAS Features	Reliability, Availability and Serviceability with shared power and cooling system design
Rack Level Power	16kW - 32kW depending on system configuration
Management	Integrated IPMI 2.0 remote management
Support	2-year warranty on parts and labor
Options	Cluster management software Linux or Windows® operating system Cluster software compilers development tools Installation and onsite maintenance services

Appro offers configuration options to include Appro Cluster Engine (ACE) Management Software or Rocks+ and MOAB. All options can be easily tested and pre-integrated as a part of a complete package to include HPC professional services and support.

Software Rolls Package Choices

Intel Cluster Ready Roll
Intel Developer Roll (Compilers)
PGI Roll (Portland Group Compilers)
MOAB Roll (Cluster Resources)
LSF Roll (Platform)
TotalView Roll (Debugger)
CUDA Roll (NVIDIA/Tesla)
Absoft Roll (Compilers)
Support Roll

Cluster Management Options

Workload Management, Cluster Management, and Access Portal benefits:

- Platform management and monitoring
- Remotely manage up to thousands of compute nodes through support for management and sub-management servers

Security via user ID/Password using SSH/SHA

Web interface option

Support for GigE and InfiniBand™ Interconnect configurations

Appro HyperPower™ Blade Clusters

(Number of Scalable Compute Nodes per Standard 42U Rack Cabinet)

	1 Rack	2 Racks	4 Racks	8 Racks	12 Racks
# of CPUs	80	160	320	640	960
# of CPU Cores	480	960	1,920	3,840	5,760
# of GPUs	80	160	320	640	960
# of GPU Cores	35,840	71,680	143,360	286,720	430,080
CPU Memory Capacity	3.84TB	7.68TB	15.36TB	30.72TB	46.08TB
GPU Memory Capacity	240GB	480GB	960GB	1,920GB	2,880GB
GPU Single Precision	82.4TF	164.8TF	329.6TF	659.2TF	988.8TF
GPU Double Precision	41.2TF	82.4TF	164.8TF	329.6TF	494.4TF

Appro HyperPower™ Building Block Configuration

GreenBlade System

Device Bays	Up to 5 CPU blades and 5 GPU expansion blades
Form Factor	5U
Power Supply	Up to four 1625W high-efficiency PSUs redundant N+1 configuration
Cooling	Up to 3 cooling fan units (CFU) Each CFU has redundant cooling fans
Ethernet I/O	On-board 2 port GbE LAN (RJ45)
Dimensions	8.75"H x 19"W x 26"D
Weight	173.6 lbs (78.4kg) max.
Management	Support for high-speed interconnects, Appro Cluster Engine (ACE) Management SW Windows® or Linux OS

GPU Expansion Blade Node

GPU	NVIDIA® Tesla™ M2050
GPU Capacity	Two
GPU Cores	Up 896 GPU cores per expansion blade
Memory Capacity	Up to 6GB ECC GDDR5 memory
Dimensions	5"(H) x 1.75"(W) x 25"(D)
Weight	~14 lbs (~6.4kg) each GPU expansion blade

CPU Blade Node

Processor	Quad/Six-Core Intel® Xeon® Processor 5600
Processor Capacity	Two
Chipset	Intel® 5520 chipset
System Bus	Intel QuickPath Interconnect (QPI)
Memory Type	Support for 800/1066/1333 MT/s ECC RDIMM DDR3 memory
Memory Capacity	Up to 96GB in 12 DIMMs across six memory channels (3 channels per processor)
Disk Controller	Intel I/O controller hub (ICH10R)
Drive Bays	Up to two fixed 2.5" SATA HDDs
Storage Capacity	1.0TB SATA
Storage Expansion	Optional storage expansion accessory with up to 4.0TB per compute blade
GPU Expansion	GPU Expansion Blade is paired with the CPU Blade node using PCIe link. Supports 2 GPUs offering 896 GPU cores per expansion blade.
Graphics	On-board VGA graphics
Network Interface	On-board dual-port Ethernet controller (optional QDR InfiniBand with QSFP connector)
Input/Output	Two USB 2.0 compliant ports, two RJ-45 LAN ports, VGA, serial, optional QSEP (B)
Expansion Slots	One x16 PCIe Gen2 PCI riser slot capable of supporting a low-profile add-in card
Power and Cooling	See GreenBlade System specifications
Weight	10.8 lbs (4.9kg) per node
Dimensions	5"Hx 1.75"W x 25"D (127 x 44.5 x 635 mm)
Temperature	Operating: 10 - 35°C, Storage: 70°C
Remote Server Mgmt	IPMI 2.0 compliant, integrated baseboard management controller (Integrated BMC)

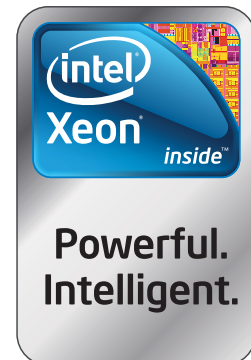


Supercomputer Solutions

Appro International, Inc.

446 S. Abbott Avenue, Milpitas, CA 95035, USA
 1.800.927.5464 (US only) • 1.408.941.8100 Main
www.appro.com

Copyright © 2010 Appro International, Inc. All Rights Reserved. Technical information in this document is subject to change without notice. Reproduction, adaptation, or translation without prior written permission is prohibited, except as allowed under the copyright laws. Intel, the Intel logo, Xeon are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.





Appro International, Inc. | 446 South Abbott Ave. | Milpitas, CA 95035, USA
1.800.927.5464 (US only) | 1.408.941.8100 Main | 1.408.941.8111 Fax
info@appro.com | www.appro.com