PLATFORMS

T-Blade V-Class Series
V205S & V205F Compute Modules
Overview

T-Blade V-Class Series
V205S & V205F Compute Modules
Overview

Rev.: 2.0, 01/05/2012

Table of Contents

T-Blade V-Class Series

V205S & V205F Compute Modules Overview

1. V205S and V205F Compute Modules

V205S and V205F are initial V-Class Series compute modules developed by T-Platforms, the leading Russian HPC solutions manufacturer, to support the latest AMD Opteron™ 6200 'Interlagos' CPUs and NVIDIA® Tesla™ M-series GPU accelerators.

The V205 modules enable a variety of configurations based on the V5000 chassis designed for research and commercial applications. The modules come in two versions, - a standard-width V205S and a double-width V205F, and are available in several standard or built-to-order configurations.

1.1 V205 Compute Module Positioning

High Core Count

A fully populated 5U V5000 enclosure hosts up to 10 V205S nodes with 320 AMD Opteron™ cores (2560 cores within 42U rack), making it an appealing building block for the higher density HPC installations. To put it in prospective, the high-end TB2-XN 7U system with 32 Intel® Xeon® nodes, created by T-Platforms in 2009, contained 384 cores and was more than two times more expensive. Furthermore, the V205 nodes may be configured with a variety of CPUs from four-core models with a fixed 3.3GHz clock rate, to 16-core models with a 2.3GHz base clock rate.

Highest Performance per Watt

Compared to the 6100 'Magny Cours' series, the newest 6200 'Interlagos' series provide 10-35% improved performance in low- and multi-threaded applications with the same TDP of 115W. New power saving technologies, such as TDP Capping, introduce granular power controls to deploy more compute capacity in electrically or thermally restricted environments.

GPU-Accelerated Computing

The V205F compute module provides support for a single NVIDIA® Tesla™ M series accelerator to deploy up to 5 accelerated nodes in a single V5000 chassis. As heterogeneous computing receives wider market acceptance with many ISVs and academic institutions optimizing their code for the CUDA ecosystem, T-Platforms is planning to further expand its GPU-based offerings to bring new levels of application acceleration and performance per watt.

Super Efficient System Memory to Run Larger Models or I/O Intensive Applications

The new V205 compute module has 50% more DIMM slots than the previous generation AMD Opteron™-based T-Blade 1.1a node. The 8-core AMD Opteron™ 6220-based system can enjoy up to 16GB of DDR3 memory per core, while the 16-core AMD Opteron 6276-based configuration handles up to 8GB of DDR3 memory per core. Moreover, the memory controller provided in the new AMD Opteron 6200 series CPUs supports all four memory channels running at 1600MHz speed, significantly increasing throughput of memory reading and writing operations.

Local Storage for Your Applications

To store temporary data, customers can choose up to two 2.5" hard drives or solid state disks with the maximum total raw disk space of 2TB for both compute module types. These are available as 0/1/10 RAID levels and diskless boot options, using internal USB drive, PXE or iSCSI protocols.

Unique System Board Design

The dual-socket system board for the V205 compute module is an original design effort of T-Platforms. The 16-DIMM system board has a reduced height to fit our 5U chassis, providing customers and partners with an opportunity for further differentiation. In addition, this system board can be supplied in a version for rack-mount skinless 'twin' servers.

V205 Compute Node Highlights

Supported CPUs and GPUs:

- *AMD Opteron™ 6100 and 6200 series (4-8-12- & 16-core versions)*
- *NVIDIA® Tesla™ M2075 & M2090 (supported in V205F only)*

Memory subsystem:

- *Up to 16 x DDR3 RDIMM ECC 1066/1333/1600 MHz modules (up to 256GB/node)*

Local storage subsystem:

- *Up to 2 x 2.5" cold-swappable SATA 2.0 (3Gb/s) drives per node*

Expansion slot:

- *1 x16 PCIe Gen 2.0 expansion slot for low profile MD2 adapter (supported in V205S only)*

Network/Interconnect ports

- *Two GbE ports and an optional integrated QDR Infiniband /10GbE VPI controller (one QSFP port)*

1.2 V5000 Chassis Positioning

V205S/F compute modules are designed for the V5000 5U chassis, which has an improved power management and centralized monitoring subsystem (Pictures 1 & 2).

The V5000 chassis has no built-in network or interconnect switches. Combined with computing modules equipped with QSFP ports, it enables users to create compute clusters containing up to 648 nodes. The use of external Infiniband edge switches is an efficient solution for scalable systems, which can also help to avoid a port oversubscription if required.

This highly functional chassis with redundant hot-swap components is positioned as a solution for a wide range of HPC users.

Picture 1. V-Class chassis front panel

Picture 2. T-Blade V chassis rear

V5000 Chassis Highlights

Form factor:

- *5U, for standard 19" rack cabinets with a depth of not less than 1070mm*

Maximum hot-swappable compute module count:

- *10 dual processor S-type compute nodes*
- *5 dual-processor F-type compute nodes with GPU*
- *Compute module mix and match functionality*

System management:

- *Integrated cold-swappable 1U module with control panel and built-in Fast-Ethernet switch equipped with two external GbE uplink ports for the consolidation of remote compute module/ chassis monitoring and management functions*
- *Support for iKVM, Remote Media, Serial over LAN, IPMI2.0 over LAN*

Cooling subsystem:

- *3 hot-swappable redundant (N+1) fan modules*

Power subsystem:

- *3 or 4 hot-swappable redundant (N+1) 1600W PSUs*
- *80Plus Platinum PSUs (94%)*

Peak power consumption:

- *4700W (preliminary information for the maximum configuration based on 5 F-type modules with 10 GPU NVIDIA® Tesla™ M2090 accelerators)*

* For more information on V5000 chassis, please see 'V5000 Chassis Overview' document.

2 V205S/F Compute Modules Details

The V5000 chassis hosts two types of AMD Opteron™-based compute modules:

- *The V205S, an x86 HPC module in a standard tray (Picture 3)*
- *The V205F, x86 node in a double width tray with GPU accelerator installed. An additional width is required to provide the necessary space and cooling for NVIDIA Tesla™ M-series GPU (Picture 4).*

Picture 3.
Standard-width V205S compute module with one SSD disk drive (top view, w/o duct cover)

Picture 4.
V205F double-width CM with NVIDIA Tesla M2090 accelerator (top view, w/o duct cover)

The V5000 chassis supports single type or mixed S & F node configurations, irrelevant of CPU platform vendor or CPU model. Customers can mix & match different types of compute nodes in a random order. In case of unpopulated compute bays, blank panels have to be used for proper system cooling. The management system has node type and presence detection logic to properly visualize each system configuration in the GUI and on the chassis control panel. Compute module hot swap is supported.

For more information on node installation, please see 'V5000 Chassis Overview' document.

2.1 V205S & V205F Compute Modules Design

Unlike most twin-class systems, both module types are supplied in closed trays with a removable top panel. To increase system reliability and uptime, the V205 compute modules do not have built-in fans and are virtually devoid of cable connections. The only exception is the NVIDIA® Tesla™ M20xx accelerator's auxiliary power cable required for V205F module.

Both CM trays have honeycomb/rectangular panel openings to provide sufficient node cooling by in-chassis cooling fan modules (CFM). Outer panels of both trays feature an I/O panel (Picture 5). The V205S module also has a bracket to install one PCIe x16 LP MD2 expansion adapter..

See details on the CM's ports and displays in section 2.2.

There is a specialized hot plug connector and a guide on the module's inner side to connect the CM with the chassis' midplane.

Picture 5.
V205S (left) and V205F (right) compute modules

Both S- and F-type trays contain the same dual socket T-Platforms server board with a CardEdge-type passive bridgeboard connector to the V5000 Chassis midplane (Picture 6). The System board comes with two SATA connectors, supporting directly attached 2,5" SATA 2.0 cold-swappable drives.

Picture 6. Bridge board in V205S.

The V205F compute module contains airflow guides for cooling the NVIDIA® Tesla™ M Series accelerator (Picture 7).

Picture 7. V205F CM with NVIDIA Tesla M2090 accelerator

2.2 V205 System Board Overview

The V205 system board (Picture 8) is based on the AMD G34 platform and supports the latest 4/8/12- and 16-core AMD Opteron™ 6200 Series and 8/12-core 6100 Series CPUs with TDP of up to 115W.

The System board contains a dedicated CardEdge connector to dispatch input power and display and control signals. To accommodate installation in a 4U bay, the V205 system board's height is a few millimeters shorter compared to most third-party G34 boards with 16 DIMM slots available on the market.

Picture 8. V205-S1A system board

The V205 system board dimensions are:

- *Length: 506mm (20");*
- *Width: 165mm (6.5")*

The V205 system board is delivered in B (blade) and S (server) SKUs with the latter featuring a dedicated Fast Ethernet management port and additional SATA and power connectors:

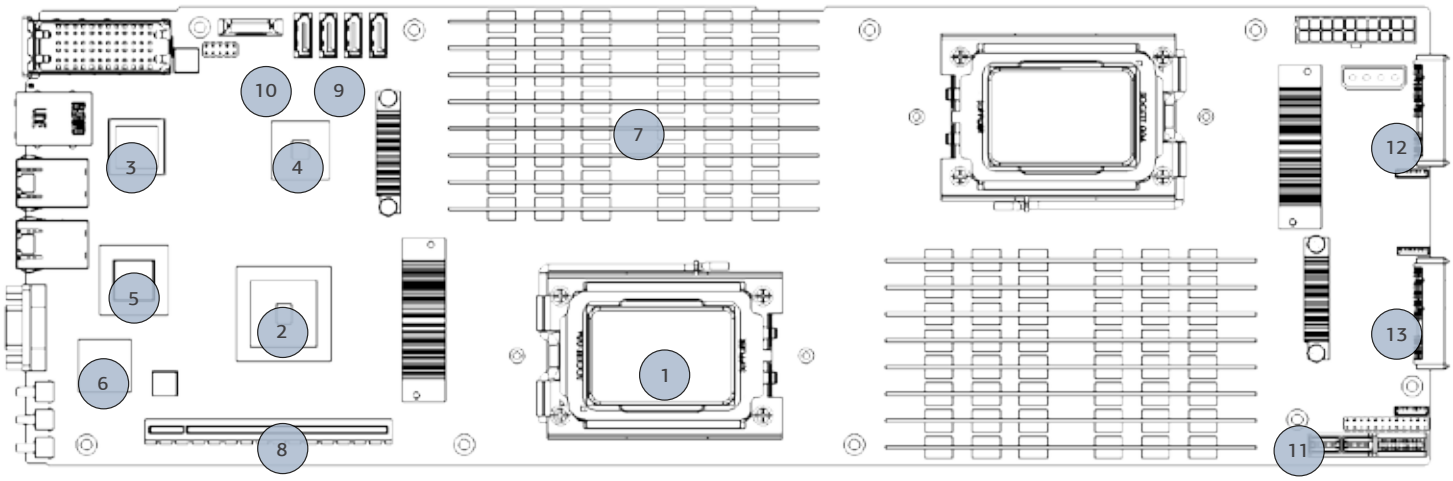
B version:

- *V205-B1A – 'Blade' SKU with integrated IB controller to install into the V5000 chassis*
- *V205-B0A – 'Blade' SKU without integrated IB controller to install into the V5000 chassis*

S version:

- *V205-S1A – Server SKU with integrated IB controller to install into 'twin' server*
- *V205-S0A – Server SKU without integrated IB controller to install into the 'twin' server*

For more information on the S SKU availability, please contact your T-Platforms representative.



Picture 9. V205 system board components layout (prototype layout with all possible options, ports, and connectors)

CPU sockets:

1. G34 socket for AMD Opteron™ CPUs with TDP up to 115W

On-board controllers:

2. AMD SR5670 North bridge
3. AMD SP5100 South bridge
4. Mellanox ConnectX®2 single port IB controller (optional)
5. Intel® 82580DP dual-port GbE controller
6. ASPEED 2050 BMC/VGA chip

Expansion slots:

7. 16 DDR3 DIMM slots (8 per socket)
8. Single PCI E x16 Gen 2 slot
9. 4 SATA 2.0 ports (optional)
10. On-board USB2.0 port

Dedicated connectors:

11. CardEdge connector for midplane
12. SATA 1 connector
13. SATA 2 connector

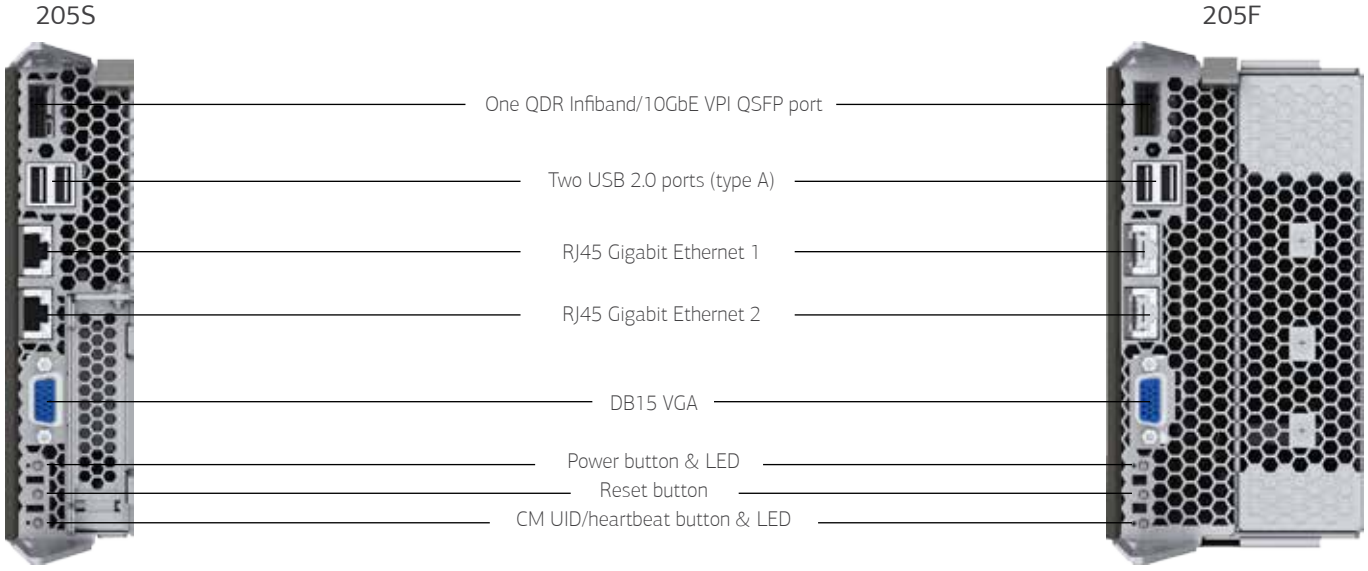


Table 1. I/O panels of V205S and V205F CMs

2.3 AMD 6000 Platform Overview

AMD Opteron™ 6000 Series CPUs use Direct Connect 2.0 architecture based on HyperTransport™ technology, as well as on-die memory controllers.

Each server processor has up to 4 point-to-point HyperTransport 3.0 links to connect sockets and I/O devices over a unified, scalable interface.

The memory controllers built in AMD Opteron™ 6200 CPUs support DDR3 memory at speeds of 1066/1333/1600MHz for dual DIMM per channel (DPC) system boards like ithe V205S/F.

AMD was the first-to-market x86 server platform with integrated on-die memory controllers and scalable point-to-point system topology, abandoning a front system bus, a feature that contributed to x86 platforms to becoming the dominant server architecture.

2.3.1 AMD Opteron™ 6100 Series CPU

Introduced early in 2010, the AMD Opteron™ 6100 Series is the world's first mass-market 12-core server CPU. Based on the Magny-Cours architecture, the 6100 Series is 2-4-socket scalable, removing the artificial cost discrepancy between 2 and 4P designs.

The CPU has a 4-channel DDR3 memory subsystem, supporting DIMM speeds of up to 1333MHz. Manufactured using a 45nm process, it cannot reach clock speeds of the Intel® Xeon® X5600 series, however, its core count, memory throughput and thermal power makes Magny-Course a great candidate for many HPC applications, which do not rely on high CPU frequency.

The highly stable AMD platform strategy protects customers' investments, supporting the entirely new Bulldozer core microarchitecture of the next generation Opteron™ 6200 processor series.

Model	Clock Rate.	L2 Cache	L3 Cache	Number of cores	HT	TDP	Socket
61KS	2.0 GHz	8 x 512KB	2x6MB	8	6400 MT/sec	115 W	G34
61QS	2.3 GHz						
6128	2.0 GHz						
6134	2.3 GHz						
6136	2.4 GHz						
6140	2.6 GHz						
6124 HE	1.8 GHz					85 W	
6128 HE	2.0 GHz						
6132 HE	2.2 GHz						
6168	1.9 GHz	12 x 512KB		12		115 W	
6172	2.1 GHz						
6174	2.2 GHz						
6176	2.3 GHz						
6164 HE	1.7 GHz					85 W	
6166 HE	1.8 GHz						

Table 2. AMD Opteron 6100 Series V205 support matrix

2.3.2 AMD Opteron™ 6200 Series CPU

The latest Interlagos chip released in November 2011 is based on the entirely new Bulldozer two-core module design. Each Interlagos chip contains two dies with several Bullzoder modules, connected via scalable HyperTransport 3.0 links.

The new 32nm chip offers impressive performance and scalability gains, along with an array of innovative features, including:

- *Various core counts of 4-8-12 and 16 cores*
- *Increased bandwidth and memory speed of up to 1600MHz per channel*
- *4 HyperTransport 3.0 links with up to 6.4GT/sec*
- *FlexFP 256 (or 2x128) bit-wide floating point unit, supporting Advanced Vector Extensions (AVX) to increase integer and floating point operations in scientific, multimedia and financial applications, and overall parallelism*
- *New TurboCore 1 operational state to increase clock rate of all cores by 500 MHz to 1 GHz*
- *TurboCore 2 mode increasing clock rate of half of cores by 1GHz*
- *Aggressive clock gating to reduce power consumption*
- *Reduced minimal P-state: 500MHz*
- *Enhanced C1E and new C6 power states with CPU cache data saved to RAM to enable CPU deep sleep modes when idle.*
- *TDP Capping technology to reduce CPU power envelope by 1-watt increments to optimize computational density or to meet datacenter power consumption limits*

2.3.3 HyperTransport 3.0 Technology

HyperTransport 3.0 technology is a high-speed (6.4GT/sec), low latency, point-to-point link protocol. HT operates both in cache-coherent inter-CPU and peripheral modes. It also facilitates more efficient and flexible use of memory in multi-processor systems, using efficient inter-processor communication arbiters. HT Assist™ reduces coherence traffic overload on the HT links, resulting in faster queries in both cache and compute-intensive applications.

Модель	Базовая частота / частота NB	Режим Turbo CORE P1	Режим Turbo CORE P0	Кэш L2	Кэш L3	Кол-во ядер	Частота HT	Мощность, TDP	Тип гнезда
6204	3,3/2	NA	NA	2x2M	16M	4	6400 MT/c	115 Вт	Socket G34
6276	2,3/2	2,6 ГГц	3,2 ГГц	8x2M	16M	16	6400 MT/c	115 Вт	
6274	2,2/2	2,5 ГГц	3,1 ГГц						
6272	2,1/2	2,4 ГГц	3 ГГц						
6238	2,6/2	2,9 ГГц	3,2 ГГц	6x2M		12			
6234	2,4/2	2,7 ГГц	3 ГГц						
6220	3,0/2	3,3 ГГц	3,6 ГГц	4x2M		8			
6212	2,6/2	2,9 ГГц	3,2 ГГц						
6262 HE	1,6/1,8	2,1 ГГц	2,9 ГГц	8x2M	16M	16	6400 MT/c	85 Вт	

Table 3. AMD Opteron 6200 Series V205 support matrix

2.3.4 AMD Chipset

The V205 system board is based on the AMD SR5670/SP5100 controllers supporting a wide range of server and desktop AMD CPUs. Produced from early 2010 on, it is a part of AMD's Stable platform initiative to protect customer investments by delivering unified socket and platform infrastructure support for next-generation CPUs.

The AMD SR5670 is PCI Express 2.0 controller supporting 30 PCI Express 2.0 lanes with HyperTransport 3.0 (5.2GT/s) interface to connect to AMD Opteron™ processors.

SR5670 controller selection was not accidental. SR5670 is the most balanced AMD solution with both adequate opportunities for I/O and low heat generation (15.4W versus 18W for 42-lane SR5690) and sufficient SATA functionality. SR5670 uses an x4 A-Link interface to connect the SP5100 peripheral controller.

The AMD SP5100 provides a 3Gb/sec interface to connect up to 6 Serial ATA devices and supports SW RAID basic functions, using the Promise RAID stack. The controller also supports up to 12 USB 2.0 ports, PCI bus 2.3 and LPC/SPI interfaces.

2.3.5 New AMD-P Energy Saving Technologies in AMD Opteron™ 6200 Series CPUs

AMD-P technology is an umbrella name that incorporates all of the power management technologies available in AMD Opteron™ 6100 and 6200 CPU series. There are three main innovations in the AMD-P technology to highlight:

- ***TDP Power Capping*** *makes it possible to set TDP power limits in 1 watt increments. First it engages aggressive clock gating schemes on 'Bulldozer' modules before it changes the processor p-state. Customers can actually buy higher performance CPUs and then modulate them down to the required TDP. This can prove important for environments where there is a need for higher computational density per each sq. meter, while being constrained by power limits set for every cabinet.*
- ***C1E*** *is a power management state that allows the processor to reduce power consumption not only by cores, but also by the memory controller and other elements. HyperTransport™ technology links may be halted too.*
- ***C6*** *is a new power state to provide efficiency when both 'Bulldozer' module cores are idle with the cache state saved to the DRAM modules. The technology allows saving up to 85% of energy by turning most of transistors off.*

2.3.6 FlexFP Technology with AVX Extensions Support

The Coprocessor microarchitecture called FlexFP is attractive with its efficiency and flexibility. Customers get 256-bit AVX support for new software, while still providing the best performance for current software, most of which isn't using AVX yet. FlexFP supports two 128-bit paths or a single 256-bit path without wasting die space and energy on register hardware that isn't being used. This reduces the CPU transistor base and power required for the rarely used features.

With the Opteron™ 6200 Series release, AMD is the first to market with a server processor supporting the Intel® Advanced Vector eXtensions (AVX) instruction set, 256-bit floating point operations, and also SSE4.1, SSE4.2, AES, CLMUL, and 128-bit SSE instruction sets, including FMAC, XOP, FMA4 and CVT16 operations.

AVX adds 12 new instructions and increases XMM register size from 128 bits to 256 bits. Using the latest compilers, users can double the number of operations executable by applications. Also, performance of SIMD instructions, involved in many scientific and multimedia applications, is accelerated.

3 DDR3 Memory Subsystem

A four-channel DDR3 memory controller integrated directly into the AMD Opteron™ chip reduces delays of inter-chip data exchange and enables users to scale bandwidth up by adding additional, or more powerful CPUs to the system.

The V205 motherboard uses single stripe memory design to reduce latencies and support higher frequency DIMMs. Online Spare RAS mode is also supported.

3.1 Registered DIMMs

Registered DIMMs support an increased amount of memory, as the memory controller manages addressing signals and commands only for the register chip to reduce the electrical load on the controller itself.

V205 compute nodes support two Registered DIMMs per each of four memory channels, providing total RAM size of 256GB per node. T-Platforms tested registered DIMMs with ECC at clock rates of up to 1600MHz (Table 4).

Memory clock rate depends on the module type, memory configuration, and CPU model. 2, 4 and 8GB ECC RDIMM modules are supported.

Customers can also use low voltage (LV) 1.35V DDR3 RDIMMs which improve nodes' thermal performance to reduce total system power consumption.

	Registered DIMMs
Clock rates	1066, 1333 and 1600MHz
Rank number	1, 2 or 4
DIMM size	1, 2, 4, 8 or 16GB*
Maximum number of DIMMs per channel	2 (for V205)
Voltage	1.5V and 1.35V
DRAM technology	x4 or x8
Temperature sensor	Yes
ECC	Yes
Advanced ECC	X4 DIMMs only
Address Parity	Yes

Table 4. V205 DDR3 memory support matrix.

	DIMM 1	DIMM 2	Max MHz, 1.5V DIMMs	Max MHz, 1.35V DIMMs	Max GB/ channel
RDIMM	1R or 2R	Empty	1600 MHz	1333 MHz	8 GB
	1R	1R	1600 MHz	1333 MHz	8 GB
	1R or 2R	2R	1333 MHz	1333 MHz	16 GB
	4R	Empty	1333 MHz	1066 MHz	16 GB
	4R	1R, 2R or 4R	1066 MHz	800 MHz	32 GB
LR-DIMM	4R	Empty	1600 MHz	1333 MHz	16 GB
	4R	4R	1333 MHz	1333 MHz	32 GB

Table 5.
Registered DDR3 DIMM population guidelines for AMD Opteron™ 6200 Series CPU-based configurations

RDIMM	1R or 2R	Empty, 1R or 2R	1333 MHz	1333 MHz	16 GB
	4R	Empty	1333 MHz	1066 MHz	16 GB
	4R	1R, 2R or 4R	1066MHz	800MHz	32GB

Table 6.
Registered DDR3 DIMM population guidelines for AMD Opteron™ 6100 Series CPU-based configurations

* At the time of this document writing, 16GB modules have not been tested

4 Using GPU-based Accelerators (V205F)

GPU-based acceleration is now a major trend in HPC, with heterogeneous systems based on NVIDIA® Tesla™ technology steadily growing their share in the TOP500 list. NVIDIA is setting the pace, proactively working with OEMs and ISVs to develop a solid ecosystem of HW, tools and applications to support GPU acceleration-based computing.

As an NVIDIA's OEM partner, T-Platforms develops and delivers heterogeneous computing systems, and offers applications optimization services for the NVIDIA CUDA ecosystem.

The V205F compute node supports one M-Series NVIDIA® Tesla™ GPU via a single PCI Express Gen 2 x16 expansion slot.

The Tesla™ 20 Series offers more than 10 times the performance of 4-core x86 CPUs in double precision operations, supporting ECC memory. Tesla™ M accelerators have an enhanced reliability and options for seamless integration with system monitoring and management tools.

Compared to the V205S, the V205F computational modules equipped with M2090 NVIDIA® Tesla™ accelerators enhance peak performance more than three times, twice improving the energy efficiency of computing, while increasing chassis computing density by 1.5.

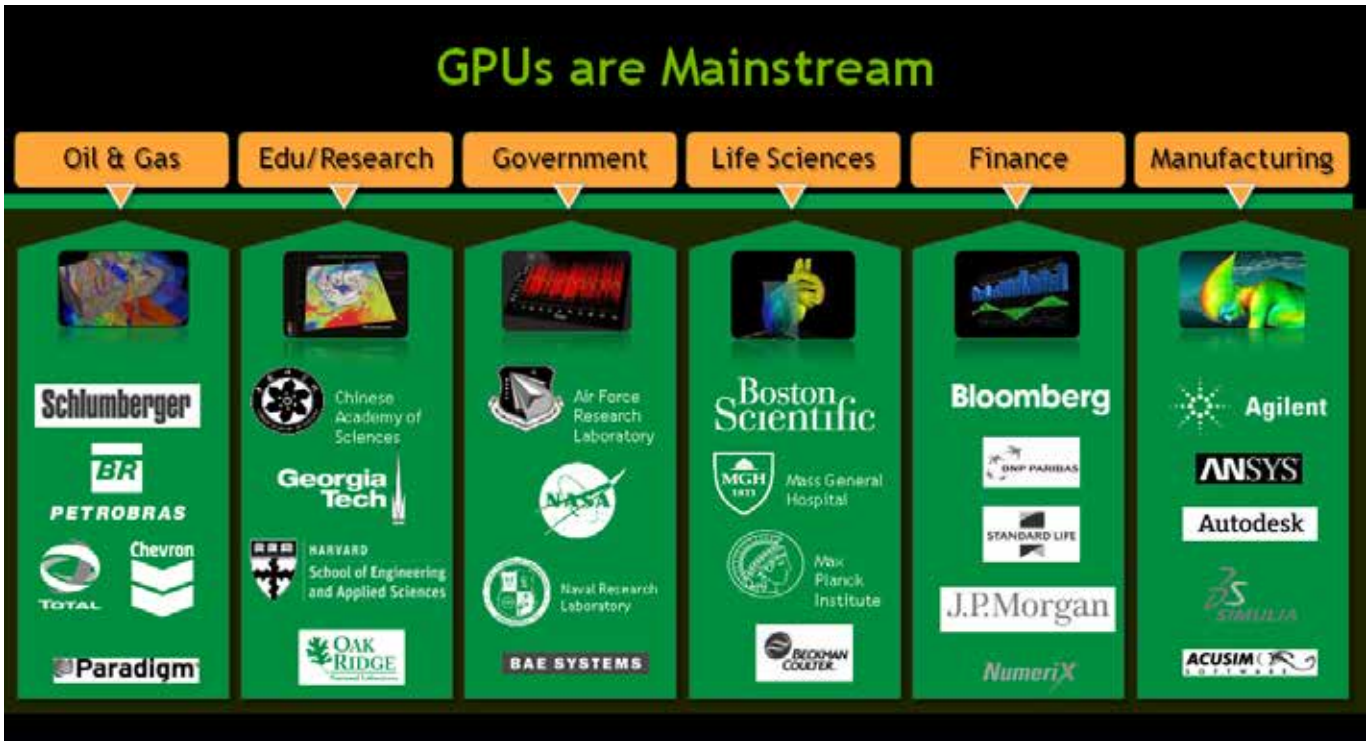
The NVIDIA® Tesla™ M Series accelerator is a double-wide, full-height PCI Express 2.0 adapter, based on NVIDIA Fermi GPU microarchitecture. It contains a GPU controller, 6 GB of high-speed GDDR5 memory and a passive heat sink cooled by a front chassis cooling modules (Picture 11).

Accelerators can be configured by the administrator in ECC-enabled mode to correct single-bit errors and detect double-bit errors, reducing available memory up to ~5.25GB. Using NVIDIA tools, the administrator also can reduce maximum allowed TDP range.

The Tesla™ M2075 accelerator is a midrange product with a TDP of 200W, while the M2090 accelerator is the most powerful product with a maximum TDP of 225W (Table 7).

When installed in V205F nodes, an accelerator is connected to the motherboard via an auxiliary 8-pin connector to provide the required power.

Picture 11. NVIDIA® Tesla™ M20x0 accelerator



Picture 10. Adoption of GPU-based computing in various vertical markets

Model	Features	Number of cards supported | Peak Performance	Board TDP
NDIDIA Tesla M2090	GPU • *Number of processor cores: 512* • *Processor core clock: 1.3 GHz* Board • *PCI Express Gen2 x16 system interface* • *Physical dimensions: 4.376 inches x 9.75 inches, dual slot* External Connectors • *None* Internal Connectors and Headers • *One 6-pin PCI Express power connector* • *One 8-pin PCI Express power connector* Memory • *Memory clock: 1.85 GHz* • *Interface: 384-bit interface* • *6 GB of GDDR5 SDRAM (5.25 GB w ECC enabled)* BIOS • *2Mbit Serial ROM*	1 card | 665GF DP	<=225W
NVIDIA Tesla M2075	GPU • *Number of processor cores: 448* • *Processor core clock: 1.15 GHz* Board • *PCI Express Gen2 x16 system interface* • *Physical dimensions: 4.376 inches x 9.75 inches, dual slot* External Connectors • *None* Internal Connectors and Headers • *One 6-pin PCI Express power connector* • *One 8-pin PCI Express power connector* Memory • *1.566 GHz clock* • *384-bit interface* • *6 GB GDDR5 SDRAM (5.25 GB w ECC enabled)* BIOS • *2Mbit Serial ROM*	1 card | 515GF DP	<=200W

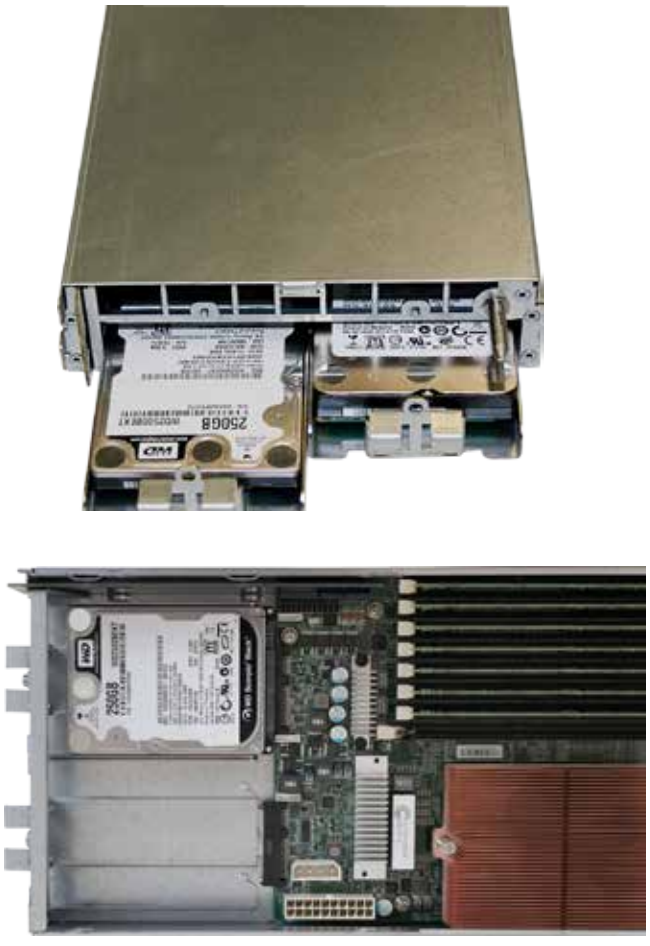
Table 7. NVIDIA® Tesla™ M Series accelerator features

5 Disk Subsystem

Both V205S and V205F type compute modules contain built-in SATA 2.0 controllers and support a maximum of two SATA disks. Customers also can order diskless node configurations bootable over iSCSI or PXE protocols, or from internal/external USB media.

AMD SP5100 South Bridge also supports Promise Software ROMB technology, implementing RAID levels 0/1/10 for 2-drive subsystem on a number of Microsoft® and Linux operating systems using Promise® RAID Option ROM and WebPAM interface.

5-, 7-, 9,5-, 12,5- and 15mm high 2,5" Hard Disk Drives and Solid State Disks are supported. Drives are installed on skids and connected directly to the motherboard without cables, using a SATA CardEdge connector. Drive cold swap is supported, by first removing the compute module from the chassis, followed by drive extraction, without the need for access to the motherboard (Figure 12).



Picture 12.
V205S node with cold swappable 2.5" drives.

Customers can decide between 7,200RPM or 10,000RPM enterprise SATA 2.0 & SATA 3.0* hard drives in different capacities, or opt for the lowest power and highest IOPS performance solid state disk technology.

Seagate Constellation® drives family for "nearline" applications are offered in 2 capacity options, featuring 1.2M-hour mean time between failures (MTBF) rating and the industry's lowest average operating HDD power of 3.11W.

The second generation, Seagate Constellation®.2 family, is the only 2.5-inch enterprise-class hard drive lineup with 1TB capacity (at the time of the system release) to store almost twice as much data in 2.5" form factor. The 1TB drive offers MTBF of 1.4M hours and operating power consumption of under 5.43W. The higher recording density contributes to marginally faster seek times, compared to the first generation of Constellation drives.

While Seagate Constellation® families are optimized for "nearline" applications delivering relatively low power consumption, up to 1TB capacity and satisfactory performance, the Western Digital VelociRaptor® drives are optimized primarily for speed. WD's SATA 2.5" drives are offered in 4 capacity options. At 10,000RPM they deliver 3.6ms random read and 4.2 ms write times, which is about twice as fast compared to the Seagate Constellation® drive families. However, performance comes at a cost of slightly higher power consumption and case temperature.

For customers limited by power, rather than budget constraints, looking for the highest IOPS storage subsystem, Intel® 320 and 510 series of Solid-State Disk Drives is a great choice to consider.

Where SATA HDDs generally deliver 75-100 IOPS at 7,200RPM and 125-150 IOPS at 10,000RPM, Intel MLC NAND based devices can reach 20,000 IOPS in reads and up to 8,000 IOPS in writes (@4KB blocks). The read-write latencies are reduced from the 3.6-8.4 millisecond level of SATA HDDs to a 65-90 nanosecond level, typical for Intel® SSD technology. Power consumption is another area of unsurpassed excellence of SSD technology.

The 25nm Intel® NAND Flash Multi Level Cell Memory-based 2.5" SSD 320 family is available in 6 capacity options, offering sustained read speeds of up to 270 MB/sec and sustained write speeds of up to 220MB/sec.

SSD 510 family (34nm Intel NAND Flash Multi Level Cell Memory) consists of two models, 120GB and 250GB. This family has a much higher bandwidth than SSD 320 series.

T-Platforms also plans to use the new Intel® SSD710 SSD family (High Endurace Technology eMLC).

Vendor	Product family	Disk Model	Disk size	Spindle Speed, RPM	Cache, MB	Buffer to host, GB/s*	Disk type	Physical height, mm
Seagate	Constellation.2	ST925061xNS	250GB	7200	64	SATA 6/3/1.5	HDD 2.5"	15
Seagate	Constellation.2	ST950062xNS	500GB	7200	64	SATA 6/3/1.5	HDD 2.5"	15
Seagate	Constellation.2	ST9100064xNS	1ТБ	7200	64	SATA 6/3/1.5	HDD 2.5"	15
Seagate	Constellation	ST9160511NS	160GB	7200	32	SATA 3/1.5	HDD 2.5"	15
Seagate	Constellation	ST9500530NS	500GB	7200	32	SATA 3/1.5	HDD 2.5"	15
WD	VelociRaptor	WD1500BLHX	150GB	10000	32	SATA 3/1.5	HDD 2.5"	15
WD	VelociRaptor	WD3000BLHX	300GB	10000	32	SATA 3/1.5	HDD 2.5"	15
WD	VelociRaptor	WD4500BLHX	450GB	10000	32	SATA 6/3/1.5	HDD 2.5"	15
WD	VelociRaptor	WD6000BLHX	600GB	10000	32	SATA 6/3/1.5	HDD 2.5"	15
Intel	SSD 320	Multiple models	80GB	NA	NA	SATA 3/1.5	SSD 2.5"	7/9.5
Intel	SSD 320	Multiple models	120GB	NA	NA	SATA 3/1.5	SSD 2.5"	7/9.5
Intel	SSD 320	Multiple models	160GB	NA	NA	SATA 3/1.5	SSD 2.5"	7/9.5
Intel	SSD 320	Multiple models	300GB	NA	NA	SATA 3/1.5	SSD 2.5"	7/9.5
Intel	SSD 320	Multiple models	600GB	NA	NA	SATA 3/1.5	SSD 2.5"	7/9.5
Intel	SSD510	SSDSC2MH120A2XX	120GB	NA	NA	SATA 6/3/1.5	SSD 2.5"	9.5
Intel	SSD510	SSDSC2MH250A2XX	250GB	NA	NA	SATA 6/3/1.5	SSD 2.5"	9.5

Table 8. A list of disk drives, available for order with V205 nodes as of November 2011.

* All SATA 3.0 HDD and SSD devices in V205 node operate in 3Gb/sec mode (due to the integrated AMD SP5100 SATA controller limitation).

6 Network Infrastructure

6.1 QDR Infiniband and 10GbE Ethernet VPI Interconnects

The V205 system board can be ordered with an optional Mellanox® ConnectX-2 VPI controller. Virtual Protocol Interface support provides QDR Infiniband or 10Gb Ethernet connectivity via a QSFP interface; the default IB mode can be altered to 'Ethernet-only' or 'Auto' mode.

The QSFP port can connect to Infiniband passive copper cables, active copper cables, and optical cables. It can also use hybrid QSFP to SFP+ cables from Mellanox.

Mellanox® ConnectX-2 Controller Specifications:

- *Virtual Protocol Interconnect (VPI)*
- *One-chip architecture*
- *Integrated SerDes controller*
- *No need to use local memory*
- *1us MPI ping latency*
- *Selectable 10, 20, or 40Gb/s Infiniband or 10GigE*
- *PCI Express 2.0 (up to 5GT/sec)*
- *CPU offload of transport operations*
- *Atomic operations*
- *16 million I/O channels support*
- *End-to-end QoS and HW-based congestion control*
- *Hardware support for I/O virtualization*
- *HW support for TCP/UDP/IP operations (stateless offload)*
- *Fibre Channel encapsulation (FCoIB or FCoE)*

Infiniband Technology Significance

In recent years, Infiniband has become widely accepted in HPC and Enterprise Datacenters and in the emerging Cloud space. Providing low-latency, high bandwidth, low CPU overhead, and Remote Direct Memory Access (RDMA), Infiniband has become the most deployed high-speed interconnect, replacing proprietary or low-performance solutions. Infiniband architecture is an industry-standard fabric, designed to provide scalability for tens of thousands of compute and storage nodes and efficient utilization of compute processing resources.

With broader compatibility, ensured by the Open Fabric Alliance, Infiniband represents a cost-effective and power-effective unified solution for oversubscribed network topologies, 3D-Torus topologies and many others.

It also features enhanced fabric consolidation and allows multiple applications (computational, management, storage) to share the same network with no performance degradations. The result is pure savings in capital and operational expenditures.

CORE-Direct™ provides CPU offloading of such collective MPI operations as broadcasting, gathering, and communication routines.

It also features enhanced fabric consolidation and allows multiple applications (computational, management, storage) to share the same network with no performance degradations. The result is pure savings in capital and operational expenditures.

Infiniband-based MPI Clusters

MPI provides communication services for distributed processes running in HPC applications. MPI is a de-facto standard and dominant communication layer model in today's parallel cluster systems, the performance of which is highly dependent on node-to-node exchange latencies. Historically, Infiniband demonstrates ultra-low latency between applications and high bandwidth coupled with low load on CPUs.

The effectiveness of the Infiniband architecture is based on sending messages over channels, a reduced number of copies of all transmitted data in the memory, and, in contrast to TCP/IP, avoiding traffic processing by OS's stack. There are several MPI implementations based on Infiniband technology to deliver these advantages.

IB Subnet Manager, along with the Subnet Administrator, provides a relatively simple interconnect setup and different topologies, allowing administrators tools for deployment, monitoring and diagnostics of Infiniband fabric.

Storage Acceleration

To ensure high performance of storage using both block and file access methods, Infiniband RDMA protocol can be used. Support for encapsulation of Fibre Channel packages over Infiniband (FCoIB) helps to provide access to high performance enterprise class storage systems. As the technology is turning mature, many leading computational sites worldwide have started deploying level 1 multi-petabyte parallel file storage subsystems utilizing QDR or FDR Infiniband infrastructure.

Upper Level Protocols

The set of upper level protocols (ULPs), provided through OFED, allows many existing applications to take advantage of Infiniband. A rich set of ULPs provided by the Open Fabrics Alliance includes:

- *MPI-MPI ULP for high-performance clusters with full support of MPI functional commands.*
- *IPoIB – IP over Infiniband Allows applications to communicate over an Infiniband network using TCP/IP messages.*
- *iSER – iSCSI Extensions for RDMA iSCSI protocol enables connection to modular SANs over standard TCP/IP-based networks to consolidate network infrastructure based on IP or to create second-tier storage.*
- *NFS-RDMA – Network File System over RDMA. NFS is a popular network file system that provides group access over the standard TCP/IP-based networks.*
- *Lustre support – Lustre is a parallel file system, often used by T-Platforms to provide compute nodes with parallel access to data. The ability to use Infiniband's Channel I/O architecture allows each node to establish an independent, secure data channel with Lustre Metadata Servers (MDS) and associated Object Storage Servers and Targets (OSS, OST).*

The set of ULPs provides a series of interfaces to access compute, storage, networking, and other services. In doing so, each of those services uses only one underlying network, Infiniband.

6.2 Gigabit Ethernet

V205 modules come integrated with Intel's 82580DB GbE Controller, - a single, compact, low power component providing two gigabit Ethernet ports on the motherboard (LOM) to support parallel cluster auxiliary/secondary networks.

The 82580DB controller uses PCI Express x 4 (PCIe v2.0; 2.5Gbps). To send and receive management packets, the controller also connects to an AST2050 BMC ASIC on V205 motherboard. GbE ports may be configured in BIOS to send management packets simultaneously with common Ethernet traffic.

Two stacked RJ45 ports support 1000BASE-T copper connections, with 1Gb/sec full-duplex, and 10/100 Mb/sec full/half duplex operation modes.

For the full feature list please refer to Intel® 82580EB/82580DB GbE Controller Feature Software Support Summary and Intel® 82580EB/82580DB Gigabit Ethernet Controller Datasheet documents on Intel's website.

6.3 Fast Ethernet management network (optional)

V205 motherboards for twin servers can be supplied with an optional dedicated Fast Ethernet port located over two external USB ports.

V-Class V205 system boards are supplied without a dedicated management port as the V5000 chassis is provided with the management switch, consolidating the node's monitoring and management through internal ports of compute modules. This dramatically reduces the number of Cat 5 cables used to manage computing modules.

7 Compute Module Level Monitoring and Control

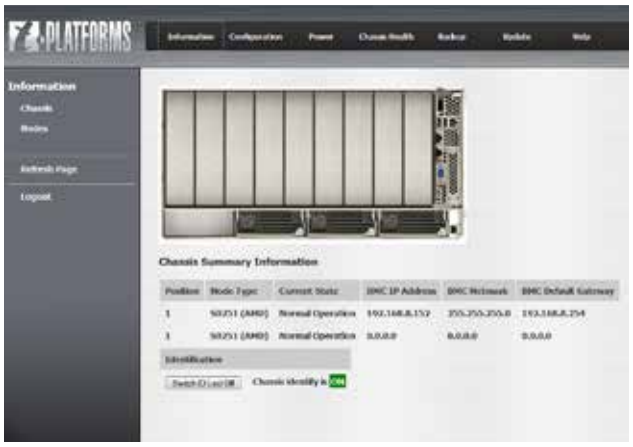
Each V205 node contains an integrated ASPEED 2050 BMC that supports the IPMI 2.0 interface, and local and remote control and monitoring via command-line and web server interfaces.

Secure access to individual computing modules can be established by direct connection to the external GbE port of each compute module, or centrally, through the single GbE port of the system management controller (SMC) in the V5000 chassis.

The IPMB/I2C bus with BMC enables a monitoring of sensors, integrated into the system board. One of the GbE ports of a compute module can be activated exclusively for service network data transfer or for mixed mode transfer of service information and standard traffic.

Supported monitoring functionality in compute modules:

- *CPU temperature*
- *North and South bridges temperature*
- *DIMM memory modules temperature*
- *Infiniband controller temperature*
- *CPU and memory voltage regulators temperature*
- *CPU core voltage*
- *Memory voltage*
- *Main voltages on the motherboard (12V, 5VSB, 3.3 V, battery)*
- *Watchdog/NMI*



Picture 13. IMU firmware interface

Supported control functionality for compute modules:

- *Allocation of static IP addresses for each BMC (through chassis' system management controller)*
- *Remote board enable, disable, cycling and reset*
- *BMC cold and hot reset*
- *Remote BIOS update*
- *Remote BMC firmware update*
- *Support for iKVM*
- *Support for Remote Media*
- *Support for Serial over LAN*

The Integrated system management controller (SMC), installed in the V5000 chassis enables centralized access to BMC's of the individual compute modules using an IMU firmware interface (Picture 13).

SMC and IMU details are available in the 'V5000 Chassis Overview' document .

8 Operating Systems Support

V205S and V205F nodes support the following operating systems:*

- *SuSE® Linux Enterprise Server 11, Service Pack 1, x86_64;*
- *Red Hat® Enterprise Linux 6 or later (6.1, 6.2) x86_64;*
- *CentOS v6.0, x86_64;*
- *Scientific Linux 6.1, x86_64;*
- *Microsoft® Windows Server 2008 R2, Service Pack 1 or later, 64-bit;*

Non-validated OSes with partial support:

- *Ubuntu latest LTS (12.04), x86_64;*
- *Debian v6.0.x, x86_64;*
- *VMWare ESX 4.0(i);*
- *Citrix XenServer 6*

9 Summary

V-Class Family

V-class is a new scale-out system for educational, academic and commercial HPC markets. Designed by T-Platforms, V-class-based cluster systems feature industry standard technologies, and are flexible enough to create just the right combination of a balanced hardware and application environment. Our innovative switchless system chassis allows customers to use interconnect and network vendors of choice, and our built-in management system provides centralized management and monitoring for chassis and nodes.

V5000 System Chassis

The V5000 System Chassis is a cost-effective chassis, addressing the performance, power and flexibility needs of many HPC users. Its functionality helps to bridge the gap between twin server platforms (featuring 2 small form factor servers in 1U rack space) and higher priced blade systems with their extended functionality, geared towards the enterprise server market.

Unlike some twin servers, V5000 does not compromise on cooling fan/PSU redundancy and hot-plug functionality, and brings centralized node manageability through the integrated Fast Ethernet management module with iKVM functionality.

The T-Blade V chassis packs a lot of computational power by hosting up to 10 two-way compute modules with external IB or GbE ports. T-Platforms also plans to release compute modules based on the Intel® Xeon® E5 2600 'SandyBridge' architecture in the first half of 2012.

The innovative power efficient design and integrated management scheme reduces cable clutter, making it a great platform for HPC tasks.

V205 Compute Nodes

V205S compute modules are attractive for use in multi-threaded environments that harness the power of 32 cores in each node. The compute module can also be configured with 8 cores at an ultra-high fixed clock rate of 3.3GHz, with increased system memory bandwidth.

The V205F compute module, based on the same system board, CPUs, memory, and disk drives, comes with the NVIDIA® Tesla™ M series accelerator.

By the second half of 2011 the AMD 6000 platform has become a critical component in half of the 25 most powerful TOP500 computer systems, providing vital competition in the standard architecture server market. Systems based on 'Interlagos' nodes increase computational density, providing users with high-performance memory, the possibility of using turbo clock rates up to 3.6GHz, features of the latest 256-bit FlexFP coprocessor supporting AVX and FMA4, and flexible power management technologies such as TDP capping.

V-Class systems based on V205 compute modules, were announced by T-Platforms in the fourth quarter of 2011. The systems are available for purchase in Russia directly from T-Platforms and via our partners in a number of other countries.

For up-to-date information on released and future systems, please visit T-Platforms website at www.t-platforms.com/products.

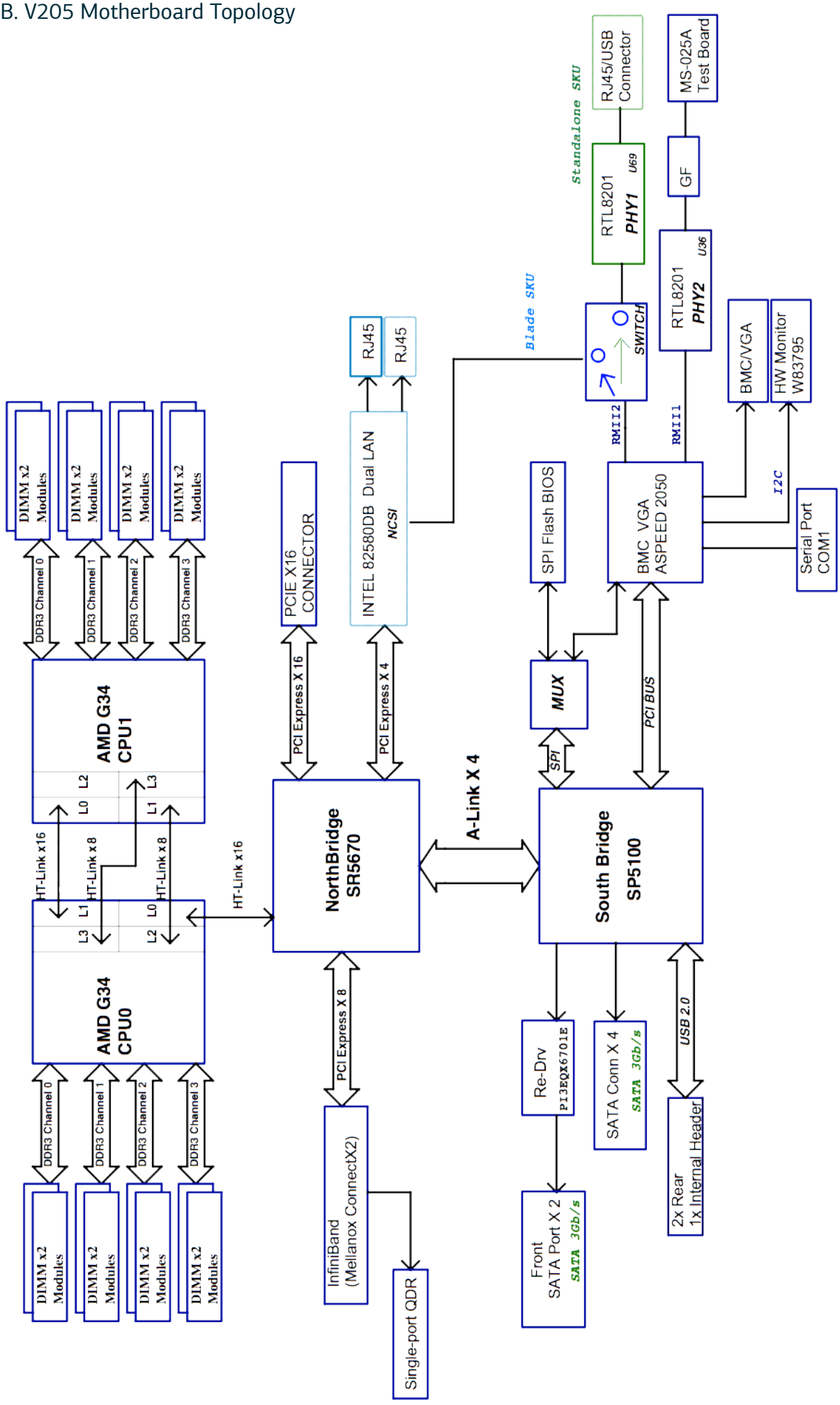
* Full support of the new microarchitecture, AMD Opteron™ 6200 Series Interlagos CPU may require an updated version of OS or a kernel patch. To obtain the current list of operating systems, please contact the company's support staff.

10 Appendix

A. V205S and V205F Compute Modules Based Systems Specification

Specifications	System based on V205S compute modules		System based on V205F compute modules	
	One node	One chassis	One node	One chassis
Chassis type	5U, for fixed rack installation			
Node type	Standard (thin)		Double-width	
Maximum hot-swappable node count	N/A	10	N/A	5
Microprocessor type, max.	AMD Opteron™ 6276 2.3GHz, 16 cores/16 threads, TDP 115W			
Maximum number of sockets/x86 cores	2/32	20/320	2/32	10/160
Support for accelerators	Not supported		NVIDIA® Tesla™ M2075 or M2090	
Maximum number of accelerators	Not supported		1	5
DP, GFLOPS peak performance	294.4	2944	959	4795
Peak performance per watt	> 0.73 GFLOPS/Watt, preliminary data		1.45 GFLOPS/Watt, preliminary data	
Supported memory type	Registered ECC DDR3 DIMM, 4 channels, 1066, 1333, 1600MHz			
Number of memory slots	16	160	16	80
Maximum RAM size	256GB	2.56TB	256TB	1.28TB
Maximum available memory per core	Up to 8GB (for 16-core CPUs)			
Storage type	Local, cold-swappable SATA 2.0 (3 GB/sec), 2.5" HDD or SSD			
Maximum number of drives	2	20	2	10
Available disk space	Up to 2TB	Up to 20TB	Up to 2TB	Up to 10TB
Compute module's chipset	AMD SR5670 controller; AMD SP5100 controller			
PCI Epress interface type	PCIe x16, Gen.2			
Number of PCI Ex16 expansion slots	1 LP MD2	10 LP MD2	1 (for GPU only)	5 (for GPU only)
Ethernet ports on computing modules	Two external GbE ports per node (Intel® 82580DB); optional 10GbE VPI® interface through QSFP port (Mellanox® ConnectX-2)			
QDR Infiniband ports on computational modules (optional)	1	10	1	5
Built-in Ethernet switch	Only built-in Fast Ethernet management network			
Built-in Infiniband switch	Not supported			
Built-in chassis/nodes control system	SMC controller with Fast Ethernet management switch and two external GbE ports; support for iKVM and Remote Media; integrated management utility			
Chassis cooling system type	Air, front-back, 3 modules with twin hot-swappable redundant (N+1) fans			
Chassis power supplies type	Three or four 1.6 kW (@220VAC) hot-swappable redundant (N+1) power supplies ('80Platinum Plus')			
Peak power consumption, W	N/A	~4200	N/A	~3300
Sustained power consumption, W	N/A	~3800	N/A	~3000
Idle power consumption, W	N/A	~1300	N/A	~1050
Power supply	1) 208-230 VAC, 50-60Hz, 4 x 8A, 1 or 3 phase 2) Some configurations also support 110-120 VAC, 50-60Hz, 4 x 16A, 1 or 3 phase			
Weight without cables and rails, kg	~5.7	~95.35	~8.10	~78.85
System dimensions with installed compute modules, mm	222.5 (5U) (H) x 443 (W) x 868 (D)			
Type of racks supported	Standard (EiA 301-D or later) 19" rack with minimum depth of 1070 mm			
Certification	TBA			

B. V205 Motherboard Topology



C. Abbreviations

- *AVX – Advanced Vector Extensions*
- *BMC – Baseboard Management Controller*
- *CUDA – Compute Unified Device Architecture*
- *DDR3 – Double Data Rate 3*
- *DIMM – Dual Inline Memory Module*
- *ECC – Error Checking and Correction*
- *EIA – Electronic Industries Alliance*
- *eMLC – enterprise Multi-Level Cell flash technology*
- *FMA – Fused Multiply Add*
- *GbE – Gigabit Ethernet*
- *GDDR – Graphics Double Data Rate*
- *GPU – Graphics Processing Unit*
- *HDD – Hard Disk Drive*
- *HS – Hot Swap*
- *IEC – International Electrotechnical Commission*
- *IPMB – Intelligent Platform Management Bus*
- *IPMI – Intelligent Platform Management Interface*
- *IPoIB – Internet Protocol over Infiniband*
- *ISV – Independent Software Vendor*
- *LOM – LAN on Motherboard*
- *MPI – Message Passing Interface*
- *NFS – Network File System*
- *OFED – Open Fabrics Enterprise Distribution*
- *QDR IB – Quad Data Rate Infiniband*
- *RAID – Redundant Array of Inexpensive Disks*
- *RDMA – Remote Direct Memory Access*
- *RU (U) – Rackmount Unit*
- *SATA (Serial ATA) – Serial Advanced Technology Attachment*
- *SIMD – Single Instruction Multiple Data*
- *SSD – Solid State Disk*
- *TDP – Thermal Design Power*
- *ULP – Upper Level Protocol*
- *VPI – Virtual Protocol Interconnect*

T-Platforms

Leninsky Prospect 113/1 Suite B-705, Moscow, Russia
Tel.: +7 (495) 956 54 90
Fax: +7 (495) 956 54 15

tPlatforms GmbH

Woehlerstrasse 42, D-30163, Hannover, Germany
Tel.: +49 (511) 203 885 40
Fax.: +49 (511) 203 885 41

www.t-platforms.com