

White Paper May 26, 2009

# Challenges and Solutions for Future Main Memory

DDR3 SDRAM is used in many computing systems today and offers data rates of up to 1600Mbps. To achieve performance levels beyond DDR3, future main memory subsystems must attain faster data rates while maintaining low power, high access efficiency and competitive cost. This whitepaper will outline some of the key challenges facing next generation main memory and the Rambus innovations that can be applied to advance the main memory roadmap.

Copyright 2009 Rambus Inc.

# 1 Introduction

Rambus.

Demand for enriched end-user an experience and increased performance in next-generation computing applications is never-ending. Driven by recent multi-core computing, virtualization and processor integration trends, the computing industry is in need of a next-generation main memory solution with anticipated data rates of up to 3200 Megabits per second (Mbps) at a similar or lower power envelope as that of DDR3-1600 solutions. The divergence of these two requirements - increased performance and reduced power - presents a difficult challenge for the design of future power-efficient memories.

In addition to power-efficiency challenges, future memory solutions face potential bottlenecks in efficiency access and capacity, both of which become more difficult to solve as memory speeds rise. For example, while memory module upgrades are the traditional way to increase capacity, the maximum number of modules allowed on a DDR3 memory channel decreases as data rates increase. In fact, due to degraded signal integrity, DDR3 memory channels can reliably support only one module at data rates beyond 1333Mbps. In addition, memory access granularity suffers as data rates increase due to the 8x to 16x multiple between the interface and core access speeds. The result is an increase to the core prefetch and a suboptimal minimum access size for future computing applications.

### Rambus Innovations Enable:

- 3200Mbps Data Rates
- 40% Lower Power
- Up to 50% Higher Throughput
- 2-4x Higher Capacity

In order to address these challenges and meet the increased performance demanded multi-core bv recent computing, virtualization and chip integration trends, Rambus has developed an architectural concept for future main memory. The architecture builds upon existing Rambus innovations and designs, such as FlexPhase<sup>™</sup> circuitry, FlexClocking<sup>™</sup> and Dynamic Point-to-Point (DPP) technologies, and introduces new concepts, which include Signaling Near Ground and Module When used in combination, Threading. these innovations can enable a future main memory system to achieve double the current data rates, with 50% higher memory access efficiency, while consuming 40% lower power than present techniques would provide.

This white paper will describe the key challenges facing future main memory, and the details and benefits of the Rambus innovations that can be applied to qo beyond DDR3 and advance the main memory roadmap. Many of these innovations patented inventions are available for licensing by Rambus.

# 2 Industry Trends and Key

Rambus.

Issues

### 2.1 Computer Architecture Trends

Fueled by the progress of Moore's law, the computational power of personal computers (PCs) has increased exponentially over the past two decades. Transistor count has increased from 275 thousand in the Intel 80386 processor to 731 million in the current Intel Core<sup>™</sup> i7 processor with quad cores. abundance This of high-performance transistors has increased computational power by nearly four orders of magnitude, from 8.5 million of instructions per second (MIPS) in 1988 to greater than 90,000MIPS in processors found in typical PCs today.

In addition to increasing clock speed, other recent trends have developed that both advance performance computing and increase associated demands on the memory system. The first is multi-core computing which incorporates several processing cores into a single chip. Some systems today incorporate 4 processors, each with 8 cores per chip, for a combined total of 32 processing cores. Each core allows for several simultaneous threads, providing an unprecedented ability to multitask applications and rapidly execute programs. Future multi-core systems are forecasted to incorporate an increasing number of processing cores for even greater performance.

A second recent trend which increases the demand on main memory performance is

virtualization, a capability that allows multiple concurrent "virtual machines", each capable of running its own independent operating system and applications simultaneously on the same hardware. A growing number of enterprise computing systems are harnessing the power of virtualization to consolidate and dynamically repurpose hardware in the data center. This in turn results in better utilization of existing hardware and reduced operating costs.

A third recent trend in advanced computing is the integration of the graphics and general purpose processing cores onto a single chip. The approach to integration varies from the combination of discrete graphics and general purpose cores, to high throughput workloads architectures where are computed in parallel to better utilize the large number of available processor cores. Throughput computing can dramatically accelerate the performance of computing systems, especially when combined with vector processing or specialized graphics cores that handle operations faster than traditional general-purpose cores.

Key Computing Trends:

- Multi-Core Computing
- Virtualization
- Converged CPU-GPU Architecture

### 2.2 Memory Data Rates

All of these recent trends advancing compute performance require an associated increase in memory bandwidth. Over the past decade, main memory data rates have doubled on average every four to five years, with memory suppliers introducing 33% faster main memory components every 18 months. Today's DDR3 main memory is shipping at typical data rates of 1066Mbps with a roadmap to 1600Mbps. Projecting from these historical trends, the next generation of main memory will support data rates up to 3200Mbps.

### 2.3 Pin Count Considerations for Total System Bandwidth

In addition to higher memory data rates, computing systems have added more parallel memory channels per processor to meet the growing demand for memory bandwidth. Starting in 1999, dual-channel memory systems were introduced in workstations subsequently and in mainstream PC platforms. The trend continues today with many processors able to interface with up to four independent memory channels.



Bandwidths shown for 64-bit memory module. Date indicates approximate industry product introduction.

Figure 1. Trends in Main Memory Data Rates

Ultimately, the number of memory channels is limited by the package pin count of the chipset or processor. Current cost-effective packages for chipsets are limited to approximately 1000 pins whereas the LGA packages used in mainstream processors are transitioning now from 776 to 1366 pins. Given this limitation, only three to four DDR3 memory channels can be implemented on these packages, even if up to 75% of the pins are allocated to the memory interfaces. Based on projected processor pin-count trends, future processors could incorporate three to six memory channels for mainstream high-end products and respectively.

Rambus.

An additional method used to improve overall performance is the integration of the memory controller with the processor. This integration increases performance by eliminating the added latency of the chipset, but can be difficult due to differences in operating voltages between the CPU and memory subsystem.

### 2.4 Power Considerations

In addition to increased performance, power is a significant issue for nearly all current computing systems. Notebooks and netbooks require reduced power to extend battery life, while desktop PCs and servers require reduced power to lower overall cost of ownership and to comply with green initiatives for reduced energy consumption.

Increased memory bandwidth requires higher data rates and a greater number of memory channels, both of which result in higher memory power. In many computing systems, the power consumption of the memory subsystem is second only to the power consumed by the processor.

The to increase performance push competes directly with the desire to reduce As performance rises, future power. memory systems are forecasted to consume an even larger portion of overall system power, as per-pin data rates and the number of memory channels increases. Computing platforms vary in system configuration and workloads. In order to reduce overall memory power in all configurations, two principal power modes of compute memory need to be addressed: Active Read/Write, and Stand-by Idle.

### 2.4.1 Active Power

When in active mode, a DRAM is actively reading or writing, and consumes power in three areas: IO signaling, system clocking, and DRAM core access. This peak power state is described in DRAM datasheets as the IDD7 power state. For example, a single unbuffered 1GB DDR3 module in active mode can consume over 3A, dissipating nearly 5W of memory power. IO signaling further adds 2W of memory power. The resulting active memory power for a single processor with four memory channels can exceed 25W, while server platforms containing up to 4 such processors can consume greater than 100W of memory power in active mode.

Despite a drop in power supply voltage from 3.3V in SDR SDRAM to 1.5V in DDR3,

active memory power has increased between successive DRAM generations. As memory bandwidth increases, the added power for higher data rates, increased rate of core accesses, and larger number of memory channels overwhelm the power savings from lower supply voltages. Reducing active memory power is therefore an important consideration for all current compute platforms.

### 2.4.2 Standby Power

When in standby mode where a DRAM is not being accessed, memory system power decreases from its peak but does not go to zero. In addition to DRAM refresh power, on-chip DLL and clock buffers on both sides of the interface must remain powered on between transactions and thus continue to draw power. Although instantaneous standby power is approximately four times lower than active power for a typical DDR3 DRAM, it remains a significant portion of the total memory power budget under typical usage models. Many computing systems are optimized to accommodate high performance under peak workloads. For systems not at peak utilization, active power decreases and standby power can become the dominant contributor to total memory power.

The problem is magnified in high capacity systems such as workstations and servers that have multiple modules in standby. For example, in a typical memory channel, only one rank within a single module is active, meaning the remaining ranks and other

### Standby power can account for up to 60% of the total memory power even under peak workloads.

modules on a memory channel are in standby mode. In high capacity servers with DDR3, standby power can account for up to 60% of the total memory power even under peak workloads. Standby power is also an issue on desktop and mobile PCs which can spend extended period at idle or lower memory utilization. As a result, reduced standby power can benefit most computing systems.



### Figure 2. Memory System Power State Example

### 2.5 Capacity

As data rates increase, signal integrity considerations can limit the number of supported on a high-speed modules channel. In previous memory memory generations, as many as 8 SDRAM modules could operate at 100Mbps on a single channel. Today only a single DDR3 module can be supported at 1333Mbps and above. With only a single module per channel, current memory systems must rely on higher density DRAMs, device stacking, or buffers to reach high capacities. All of these alternatives add cost compared to the conventional multiple module approach.





At high speed operation, the number of DDR3 memory modules per channel is limited because the electrical signaling between the DRAM and the memory controller is degraded with the higher number of electrical loads. Above 1333Mbps data rates, most main memory systems cannot natively support (without the use of buffers) module upgradeability or multiple modules. A technology to enable multiple modules at high data rates would be valuable to future main memory systems.

### 2.6 Memory Requirements Beyond DDR3

Beyond DDR3, future main memory systems are forecasted to double bandwidth while maintaining the conventional module form-factor. To achieve this bandwidth within a 64bit-wide module, data rates must reach 3200Mbps per-pin. As described earlier, these faster memory systems should ideally reduce power consumption compared to current DDR3 memory, since thermal dissipation and battery life are already challenging issues in today's systems. Future memory systems should optimize power for both active mode, where memory activity is at its peak, and standby operating modes, which dominate power dissipation of large capacity servers and mobile systems at lower utilization.

# Future Memory Requirements:

- Double the Bandwidth
- Optimize Active and Standby Power
- Multi-Module Support
- Increased Access
  Efficiency

Future Memory System Parameter	Extrapolated Requirements
Module Bandwidth	12.8-25.6GByte/second
Data Rates (Per Pin)	1600-3200Mbps
Maximum # of Memory Channels Per Processor	3-6
Active Module Power (8 device unbuffered)	<7W
Standby Idle Module Power (8 device unbuffered)	<0.8W
Sustained Transfer Granularity	64Bytes
Module upgrades	Yes
# Drops per Channel	>1
Controller Compatibility to DDR3	Optional

### Table 1. Future Memory Requirements

Beyond the physical data rates and power consumption requirements, future main memory systems will need to be highly efficient with multi-core workloads. Graphics, virtualization, and throughput computing typically access large data sets that are difficult to cache due to low spatial locality. In particular, graphics processing requires both high bandwidth and small transfer sizes which are not efficient on main memory DRAMs with large core-prefetch ratios. Increasing the cache line size can improve memory transfer efficiency but compromises processor performance due to mismatch with the line-size of lower level caches on the CPU. In addition, smaller line sizes can improve cache efficiency since more pieces of relevant data can be stored. Future main memory systems should operate efficiently with small transfers of 64byte.

Lastly, it is valuable to support module upgrades and multiple modules per channel in future memory systems. Conventional signaling can only support a single module above 1333Mbps, limiting options for higher capacities and requiring pre-installed modules to be removed during memory upgrades. This increases the costs of memory upgrades, especially for enterprise servers and workstations that may have higher capacity modules already installed.

In summary, future memory systems will need higher bandwidth with higher data rates, lower power, higher throughput and higher capacity. A higher performance, more power-efficient memory architecture is required to advance the industry roadmap beyond DDR3.

### Rambus.

### 3 Future Main Memory Architectural Concept

In consideration of these requirements, Rambus has developed an architectural concept for future main memory that will address the key issues described earlier. The architecture includes innovations and techniques to improve performance, power efficiency, memory access efficiency and system capacity.

In order to achieve this, the memory architecture incorporates a clock forwarding architecture using Rambus' FlexClocking<sup>™</sup> technology for high data rates with low standby power. Command, address, and their associated clock are routed to the DRAM devices in a topology similar to current XDR<sup>™</sup> and DDR3 memory systems. Memory transactions to each device are deskewed by the controller using FlexPhase timing adjustments. In place of data strobes, the memory controller forwards a Data Clock signal to the DRAMs and receives a Timing Reference Signal (TRS) from each DRAM during read and write operations to improve data capture. This clocking architecture can be implemented with optional controller-side compatibility to current DDR3 DRAMs.

The architecture also employs Dynamic Point-to-Point (DPP) technology. DPP technology maintains а point-to-point signaling topology for high data rates while enabling module upgradeability and increased capacity natively supported by a memory channel. To lower active IO power, the architecture uses Near Ground signaling which substantially lowers the off-chip signaling voltage. Near Ground Signaling also reduces the cost and complexity of integrating a memory controller onto a advanced processor on process technologies (45nm and below). Threading techniques such as threaded modules, dual channel modules, and device-level micro-



### Figure 4. Future Main Memory System Block Diagram

threading further reduce active core power, and improve memory efficiency with small transfer sizes that may be limited by DRAM core parameters such as tFAW and tRRD. Micro-threading on the DRAM core devices also relaxes DRAM core requirements on tCC and can improve DRAM yield.

### 3.1 Clocking Architecture

Rambus.

The clocking architecture incorporates Rambus' FlexClocking innovations and other circuit updates to ensure high performance at low power. The clocking topology also supports optional backwards compatibility to DDR3.

### 3.1.1 Clocking Key Features

The key clocking feature is very low standby current with fast turn-on time. The FlexClocking<sup>™</sup> architecture enables precise data alignment without the use of a DLL or PLL on the DRAM device. DLL or PLL circuits are typically utilized on highperformance DRAMs to minimize the timing variation of both the data transmission and receive capture points relative to the global clock. The disadvantage of these circuits is they consume power even when no data transactions are active between the memory controller and the DRAM. In recent server platforms populated with hundreds of DRAM devices, standby power can represent up to 60% of total memory power. The FlexClocking architecture adjusts for timing variability between clock and data signals, and does not consume standby clocking power between data transactions.

### 3.1.1.1 Fast Turn-On for Low-Power

To minimize power when the DRAM is not actively transmitting or receiving data, the controller transmits high-quality, а differential clock called "Data Clock" to the DRAM module. It is routed in a topology similar to the Command/Control/Address (CCA) bus and its associated clock. The Data Clock and TRS signals are active only during data transactions and are enabled with fast turn-on times. The Data Clock oscillates at the Nyquist frequency of the data rate, e.g. 1.6GHz for 3.2Gbps doubledata rate transmission. When combined with the FlexClocking<sup>™</sup> architecture, the use of a separate data clock signal running at the Nyquist rate, rather than the use of the C/A clock, can enable high data rates without a DLL or PLL on the DRAM.

On the DRAM device, the clock buffer circuit which captures this clock and distributes it to the Input Output (I/O) bit-slices must have very low timing variation due to powersupply noise. Also known as very low Power-Supply Induced Jitter (PSIJ), power supply noise can be as high as 30 to 50mV on a DRAM device even with excellent isolation and bypassing. Clock buffering with a differential Current-Mode Logic (CML) buffer improves immunity to power supply noise by distributing the clock signal differentially across the DRAM PHY. While a CML buffer consumes about 60% more power than a CMOS buffer for the same fanout, the CML buffer needs only be activated when data transactions are active, and can be disabled by a chip-select signal, an

Extended Mode Register command, or another signal sent by the controller to the active DRAM module.

Rambus.

A representative topology of a future main memory module system is shown in figure 5.

future main memory DRAM devices could implement an optional mode to output a read capture signal from the TRS pin to the memory controller on read operations in the same manner that Data Strobe is used in DDR3 today.



Figure 5. Future Main Memory Module System Topology

### 3.1.1.2 Optional DDR3 Backwards Compatibility

Backwards compatibility with DDR3 memory devices can provide added flexibility to a future memory system. Specifically, first implementations of future main memory controllers may potentially interface with DDR3 memory devices at the crossover data rates of 1600 to 2000Mbps. To simplify the design of such a memory controller,

### 3.1.2 Read-Operation Timing Evolution

Per-pin calibration can help achieve high data rates by compensating for package, system and device timing mismatches. At 3Gbps or above, per-pin calibrated data recovery in the memory controller is typically easier to implement than per-pin calibrated strobe delay. A highly accurate clock-based reference is easier to distribute and control across a 5mm-wide, 64bit interface than a highly accurate voltage or current-source reference. With per-pin calibration, the DRAM's TRS is no longer required to be a strobe signal, but can carry other useful information such as EDC for high-reliability systems. This signal can also be periodically edge-tracked by the memory controller to maintain the calibrated timing integrity.

Rambus.

Both the direct strobe sampling mode and future main memory mode are shown in the next two figures. This "bimodality" is implemented with very low overhead to the DRAM device, with the addition of minor logic gates in the simplest embodiment or up to several hundred logic gates for full EDC support.

### 3.1.3 Strobe-less Sampling Jitter Budget

An eye diagram for the passive-channel signal integrity simulations of the proposed future main memory system is shown in figure 9. The metric of success for these simulations is timing fuzz of no worse than 0.25 of a bit-time Unit Interval (UI) at the zero crossing, and at least +/- 100mV of signal swing at +/- 0.25 UI offset from the center of the eye.









These simulations include the elements of the passive channel including: substrate and connector models, input capacitance, crosstalk and ISI terms. SSO, active circuit random-jitter and power-supply induced jitter are not yet taken into account.

Rambus.

A representation of the simulated 1 Dual In-Line Memory Module (DIMM), 2-rank future main memory system is shown in figure 8. The simulated memory channel has been optimized to reduce signal cross-talk, simultaneous switching noise, and impedance mismatches.

### 3.1.4 Clocking Architecture Summary

The clocking of the main memory architectural concept includes several updates to develop a roadmap to poweroptimized, high-performance memory. This first update is the addition of a "Data Clock" signal to the C/A/Clock signal group, which is active only during data transactions anddoes not require a PLL or DLL to generate its own bit-rate clocking reference from the bit-rate command-clock signal on the DRAM. The FlexClocking architecture adjusts for timing variability between the clock and data signals, and can be dynamically updated with the TRS signal. In conjunction with this new Data Clock signal, a low-PSIJ clock buffer circuit on the DRAM can realize the timing requirements for highspeed future main memory and provide optional backwards compatibility with DDR3.



Figure 9. Passive Channel Simulation Eye Diagram of System in Figure 8





### 3.2 Breaking Capacity Barriers

Rambus.

Conventional Stub Series Terminated Logic (SSTL) memory buses in personal computers and workstations utilize a "stubbus," multi-drop topology, which allows more than one memory module per electrical trace. This topology allows for easy memory-capacity upgrades by allowing multiple modules to be plugged into the bus as needed providing benefits to system manufacturers and end users. However, as data rates increase, the multi-drop topology suffers from poor signal integrity, reducing the speed at which the memory bus can run. In multi-drop topologies, one factor that determines the speed of the memory bus is the worst-case loading characteristics in which all connectors are populated with memory modules. As a consequence, the

Beyond the DDR3 generation of memory, bus speeds will increase to the point that having more that one module per memory will become very difficult.

number of modules that can be supported in a multi-drop architecture decreases with increasing bus speed. This limitation has the effect of reducing the total memory capacity of the system. Beyond the DDR3 generation of main memory, bus speeds will increase to the point that having more than one module per memory port will be very difficult. As such, alternative methods for achieving high capacity are needed.





In very high-speed memory systems, such as in graphics applications, a point-to-point topology is typically used. In this topology, the signal trace is a transmission line with minimal discontinuities, allowing better signal integrity properties and permitting higher bus speeds. However, point-to-point systems cannot be readily upgraded because they do not allow multiple modules. The ability to increase memory system capacity by adding memory modules is such an important feature in computer systems today, that traditional main memory systems support multi-drop topologies instead of point-to-point topologies.

Rambus' DPP technology allows multiple memory modules to share the same port from the memory controller while maintaining a point-to-point topology. This allows for high-capacity memory topologies and module-based capacity expansion without significantly degrading signal integrity.

### 3.2.1 Dynamic Point-to-Point

### 3.2.1.1 DPP Overview

Rambus' DPP is a new signaling topology for increasing memory capacity at high data rates. It enables high-performance computer systems to add modules without impacting memory system performance while maintaining backwards-compatibility with most standard signaling such as SSTL, RSL, etc.

DPP uses a technique that dynamically configures the memory channel so that all

data connections are "point-to-point". With DPP-enabled systems, a portion of the memory bus is routed to each of the memory channel's DIMM sockets. Depending on how many of the sockets are populated with a DRAM module, the memory controller configures the width of the DRAM devices on each memory module. DIMM sockets without DRAM modules are populated with passive "continuity modules" to maintain the point-to-point topology. The following figures illustrate examples of DPP techniques in more detail.

Figure 11 depicts a conventional multi-drop memory bus. Here, a memory controller is connected to two module ranks. In this configuration, the DQ terminals of each module rank connect to both the memory controller and the respective terminals of the other rank through a common signal bus.

At high bus frequencies, the impedance discontinuities created by the module stubs results in a degradation of the signal integrity which limits the scalability of this architecture. As a consequence, the number of module ranks that can be supported on the multi-drop bus decreases with each successive memory generation, effectively resulting in a decrease in memory capacity for the system.



Figure 11. Dual Rank Multi-Drop Memory Bus Configuration

## Rambus.

In Figures 12 and 13, logical representations of DPP architecture examples are shown. Figure 12 illustrates a DPP architecture when a single active memory module rank is used in a dual rank expansion system. In this configuration, a low-cost, passive continuity module populates the second, unused module socket. The continuity module is used to maintain the point-to-point connectivity for half of the signal traces between the active module and the memory controller. The memory capacity of the system is defined by the bit capacity of the active rank. When a second active module rank is added to the configuration, as shown in Figure 13, the second module provides the data for the second half of the data lanes to the memory controller. The second active module rank replaces the continuity module shown in Figure 12. With the addition of the second module rank, the active modules are reconfigured to each have twice the number of addressable locations and one-half the number of DQs as compared to the nonconfigured modules. The total memory capacity of the system has been doubled, without compromising the high-speed, pointto-point signaling topology.



Figure 13. DPP Fully Populated Memory System

## Rambus.

### 3.2.1.2 DPP Channel Signal Integrity

Figures 14 and 15 show channel simulation results for both a standard multi-drop topology and a DPP topology. Each eye diagram shows the worst case configuration for a system topology supporting two dualrank modules. For the multi-drop topology showing in Figure 14, the worst-case signal integrity is observed in a system with both modules loaded. In contrast, the worst-case signaling in a DPP configuration is when only one module is loaded and the signals must cross a continuity module, adding extra connector crossings. With two modules loaded in a DPP configuration, the signaling improves over the case illustrated since the electrical path is shortened and the extra connector crossing is eliminated.

As can be seen in Figure 14, the signaling eye of the multi-drop topology does not allow enough voltage-timing margin to successfully transmit data at 3.2Gbps, even on a memory channel optimized to reduce signal cross-talk, simultaneous switching noise, and impedance mismatches. However, both the timing and voltage margin increases significantly for the DPP configuration vs. the multi-drop. With DPP, configuration can dual-module be а supported at a data rate of 3.2Gbps.



Figure 14. DQ Multi-drop Channel Eye Diagram (Dual 2-Rank Modules)



Figure 15. DPP DQ Channel Eye Diagram (Single 2-Rank Module Plus Continuity Module

Challenges and Solutions for Future Main Memory

### 3.2.2 Special Considerations

Rambus

### 3.2.2.1 Sub-Page Activation

In most DRAM designs, the page size of a single DRAM is fixed (generally to 1 to 2KB.) As modules are added in a DPP system, the number of DRAM devices in each rank of memory increases. (This is not an issue if module threading is employed. See section 3.4 for more information on this feature.) The system page size increases as modules are added resulting in more power consumption for each activation command.

Sub-page activation can alleviate this added power. With sub-page activation, only onehalf of the page is activated when the DRAM is configured to half-width. Similarly, onequarter of the page is activated in quarterwidth mode. One or two row address bits are added to the command structure to select which sub-page to activate. Correspondingly, one or two column address bits are needed to address the

appropriate fraction of the data on column accesses. In this way, the overall system page size and column granularity is maintained regardless of the number of modules in the system.

### 3.2.2.2 Module Threading

DPP topologies are synergistic with modulethreading, a concept explained in detail in section 3.4. As modules are added to a memory system, it is beneficial to treat each module as one or more independent threads. Maintaining memory threads within a module boundary allows for each module to have a different capacity, organization, or performance level without interfering with the operation of other modules.

#### 3.2.2.3 Four-Module Topology

The DPP concept can be extended to a four-module topology with the same principles explained above for the two-module case. Figure 16 depicts the basic routing topology and shows the connectivity



needed to enable the four-module system. Again, as in the dual-rank scenario, the total number of signal traces routed to each module is proportional to the number of memory module sockets. In this case 16bits are routed directly to each of the four module sockets. Where the dual-rank scenario required two sets of signal traces that could function independently, the fourrank system requires four sets of signal traces. With a single DRAM module installed in this topology, two sets of data bits require a single continuity module crossing and one set requires two crossings. Note that a single continuity module design will still meet the needs of this topology, regardless of the slot in which it is utilized.

Rambus.

Figure 17 illustrates the four-module topology with two DRAM modules installed. In this case, the DRAMs from Figure 16 have been reconfigured to one-half the original data width (x4 vs. x8.) Note also that this case is very similar to having two two-module DPP topologies on the same 64bit memory interface.

In Figure 18, the four-rank system is shown again, but in this scenario, it is fully populated with four memory modules. In this system, each of the module ranks is

Modules Installed

configured by the memory controller to have one quarter of the DQ data lanes as compared to a non-configured memory module. Each module rank provides one quarter of the total memory capacity sliced vertically.

### 3.2.2.4 DPP Command and Address

As signaling rates increase, the Command and Address (C/A) channel for modulebased memory systems also suffers from challenges with signal integrity. While data signaling to a single module is reasonable, using a shared C/A bus to control multiple modules may not be feasible. Traditional solutions for this problem include (1) using registered modules to simplify the channel and reduce the capacitive loading or (2) using a dedicated C/A bus for each module. The use of registers results in added cost and latency to the memory subsystem, while systems with dedicated C/A buses per module have unutilized pins and bandwidth in configurations where not all modules are loaded.

The dynamic point-to-point concept can be extended to the C/A bus to alleviate these issues. Figures 19 and 20 illustrate an example of DPP C/A applied to a twomodule system. Two C/A buses are routed



Figure 18. Fully Populated Four-Module DPP System

in such a way that each addresses one-half of each of the modules in the system. In this way, the worst case loading of each C/A channel is the same as it would be for a full module, with the addition of two connector crossings to support the inter-module routing. However, as illustrated in the single-module case in Figure 20, both C/A channels are still utilized when only one module is loaded, each addressing a portion of the module in place. A continuity module makes electrical connection of C/A2 from the first module slot to the second and terminates the signals from C/A1.





## Figure 20. DPP C/A with a Single Module

### 3.2.2.5 DPP Command and Address Channel Signal Integrity

As can be seen in the eye diagram in Figure 21, the signaling on the C/A channel is not significantly degraded from the additional connector crossings and routing needed to support the DPP feature at the data rates required for future main memory systems.



Figure 21. DPP C/A Channel Eye Diagram (Dual 2-Rank Module)

### 3.2.3 DPP Summary

DPP high-performance signaling is a topology that can be used both with industry standard memory modules with programmable-width DRAM devices or buffers. The DPP system permits both module expandability and high data rates since the DPP topology eliminates stub traces and the addition of expansion memory modules does not introduce additional load capacitance on the DQ data signal bus.

### 3.3 Near Ground Signaling

Rambus.

In order to reduce total memory system power, three major areas of power consumption must be improved: system clocking power, IO signaling power and DRAM core access power. In Section 3.1, a clocking architecture that incorporated Rambus' FlexClocking innovation was described as a means of reducing the system clocking power. Near Ground Signaling technology enables high data rates at significantly reduced IO signaling design power and complexity while maintaining excellent signal integrity.

There is a substantial difference in the silicon performance available on a state-ofthe-art ASIC process compared with a stateof-the-art DRAM process (e.g., the fanoutof-four delay of a 60nm DRAM process is about 400% slower than a 45nm foundry ASIC process). That performance difference has led to a large IO voltage disparity. Specifically, while a DRAM process may require 1.2V, or even 1.5V, to reliably achieve a target data rate of 3.2Gbps, an ASIC process needs only 1.0V or less to comfortably achieve the same data rate. In most cases, it is less power-efficient for a 1.0V ASIC process (e.g., 45nm and below) to interface with a 1.5V-based signal



Figure 22. Near Ground Signaling

interface than with a 1.0V-based signal interface. In this case thick-oxide transistors or stacked transistor output stages must be used, both of which are less power-efficient than the available thin-oxide devices.

Rambus.

This discrepancy in voltage and resulting power inefficiency can be avoided by implementing Near Ground Signaling. An illustration of the proposed signaling is shown in Figure 22.

### 3.3.1 Power and Cost Advantages

With Near Ground Signaling, the memory controller's IO voltage (VDD<sub>IO</sub>) can be well below the maximum voltage of thin-oxide devices, typically 0.9V to 1.0V for 45nm or below process technologies. Meanwhile, the DRAM on the other side of the channel can continue to utilize the higher voltage (i.e., 1.2V to 1.35V) its on-chip circuits require to achieve reliable 3.2Gbps signaling. previously demonstrated Rambus has reliable differential signaling at near ground levels at data rates up to 6.4Gbps. Rambus innovations can also be applied to enable robust single-ended signaling at near ground levels

Similar to pseudo-open-drain logic (PODL), the termination in Near Ground Signaling is to one of the power supplies. Unlike PODL, however, Near Ground Signaling terminates to the GND rail (0V). As the GND rail is typically the lowest-impedance plane in a memory channel, this termination scheme can also help reduce SSO noise, which is one of the performance limiters to high-rate data signaling.

Near Ground Signaling has a reduced signal swing compared to traditional SSTL or PODL, which substantially lowers IO power on both sides of the link as compared to DDR3. Beyond lower signaling power, the lower signal swing reduces the size of the IO driver circuit, enabling even further power reduction in the pre-driver and clockdistribution circuits. Simulation results indicate a savings of about 1.25mW (20 percent) per IO bit-slice, which adds up to even larger savings if the controller has three or even four 64bit memory interfaces. From a cost perspective, Near Ground Signaling also eliminates the need for thickoxide transistors on the memory controller, potentially saving at least two additional masks and two additional semiconductor processing steps. Finally, for optional backwards compatibility to DDR3 during the crossover period, a bimodal IO circuit which interfaces with low-voltage DDR3 devices can be easily implemented.

Near Ground Signaling has a reduced signal swing compared to traditional SSTL or PODL, which substantially lowers IO power.

A summary of the power simulation results is shown in Table 2 for a memory controller and DRAM on 40nm and 65nm process technologies respectively.

	VDD <sub>io</sub>	VDD (Controller/ DRAM)	Power per IO
SSTL-1.5	1.5V	0.9/1.5V	22.8mW
SSTL-1.2	1.2V	0.9/1.2V	14.7mW
Near Ground Signaling	0.5V	0.9/1.2V	1.9mW

### Table 2. Power Simulation Results

A drawback of Near Ground Signaling is the level-shifting circuits required in the receive path of the DRAM. Fortunately, high-speed, common-gate NFET level shifters which consume only about 1mW per IO can be built, and these level shifters can be disabled when there are no active data transactions.

### 3.3.2 Near Ground Signaling Summary

As shown above, there are substantial power benefits realized by transitioning future main memory from conventional SSTL to Near Ground Signaling. This is readily accomplished by terminating the IO signals to GND, such that no IO current is consumed while transmitting one of the logic states. As an additional benefit, the lower IO voltage is better matched to the operating voltage of advanced CPU's and GPU's and fits comfortably within the maximum voltage allowed for thin-oxide transistors in 45nm processes and below. This reduces the cost and complexity of integrating the memory controller on the processor chip.

### 3.4 Threading

As mentioned earlier, future memory systems beyond DDR3 call for reduced power and high efficiency with multi-core workloads. Having already addressed system clocking power and IO signaling power, threading techniques such as module-threading and micro-threading can reduce the power of memory core accesses and improve data transfer efficiency.

### 3.4.1.1 Compute Workload Trends

Modern computing systems typically employ a memory hierarchy where data for internal computational registers are supplied through nested levels of memory caches. Data requests to access cache lines are made recursively upon cache misses starting from the first-level cache through the next-level cache and ultimately to the external memory system. The data requests are typically in 64Byte chunks for modern processors which generally correspond to the line size of the first or second-level caches. The cache line size is selected through a compromise between fewer number of memory fetches (which favor larger cache lines) and maintaining more pieces of relevant data that may not exhibit spatial locality (which favors smaller cache line sizes). Additionally, the line size of the third level cache typically may be limited by the minimum efficient transfer size of the memory subsystem.

The recent trend to multi-core processing and converged graphics-compute processors is increasing the performance DRAM memory requirements on the subsystems. Multi-thread computing and graphics not only need higher memory bandwidth but also generate more random accesses to smaller pieces of data. In particular, graphics processing tends to access many small pieces of data representing polygons or blocks of a visual frame.

Rambus.



Figure 23. Memory-Cache Hierarchy of Typical Compute Systems

### 3.4.1.2 DRAM Memory Access Granularity

Small transfers of data are becoming increasingly difficult with each DRAM generation. Although the memory interface has become faster, up to 1600MHz data rates for DDR3, the frequency of the main memory core has remained between 100-200MHz. As a result, DRAMs implement core prefetch where a larger amount of data is sensed from the memory core and then serialized to a faster off-chip interface. This core-prefetch ratio has increased from 2:1 for DDR DRAMs to 8:1 in current DDR3 DRAMs. Extrapolating forward, a future memory DRAM is forecasted to implement a core-prefetch ratio such as 16:1.

The efficient burst length of each transfer from a compute memory system is determined by the core-prefetch ratio. For a DDR3 DRAM component, the minimum



Increasing DRAM Internal Prefetch

efficient transfer can be calculated by multiplying a burst-length of 8bits through an 8 or 16bit interface for a net transfer of 8-16Bytes per component. For a future main memory DRAM with a burst length of 16bits, the minimum efficient transfer per component is 16-32Bytes per component.

DDR3 DRAMs implement the burst-chop feature for access to smaller transfer sizes. However, this feature only truncates the transfer of data from a burst of 8bits to 4bits. The internal timing restrictions of the DRAM components, such as the minimum Column CAS to CAS command delay (tCCD), must still meet a minimum of 4 clock cycles. Therefore, there is no improvement in the rate of memory access and thus no gain in effective data transferred between burstlength of 8 or 4bits.

### 3.4.2 Threading Architecture

### 3.4.2.1 Traditional Single-Channel Module

The memory subsystem today's in computing platforms typically are implemented with DIMMs that have a 64bitwide data bus and 28bit а command/address/clock bus. An example module configuration places eight (x8) DDR3 components assembled in parallel onto a module printed circuit board. On a standard DDR3 DIMM module, all the devices within a module rank are accessed simultaneously with a single C/A. The resulting minimum efficient data transfer is calculated as burst length of 8 multiplied by 64bits, or 64Bytes per transfer.



### Multi-Core Processor

The minimum efficient transfer size of a future DIMM module with a 64bit bus and a burst-length of 16 is 128Bytes per access, or double the transfer size of a DDR3 module access. This larger transfer size can lead to computing inefficiency, especially on multi-threaded and graphics workloads with the need for increased access rate to smaller pieces of data.

### 3.4.2.2 Dual-Channel Modules

Rambus.

Greater efficiency with a multi-threaded workload and smaller transfers can be achieved by partitioning the module into two separate memory channels. The multichannel DIMM approach was pioneered on the dual channel RIMM 4200 module, first demonstrated by Rambus in 2002. In a dual-channel module, each side of the module implements its own command and address port, thereby reducing the minimum transfer size to one-half the amount of a standard single-channel module.

A dual-channel module can be implemented using standard PCBs and does not require additional DRAM devices. This approach implements an additional C/A port and effectively doubles the C/A bandwidth of the module. The dual-channel approach also improves the electrical performance and lowers attenuation of the C/A signal since the controller only needs to drive the C/A signal to one-half the number of devices compared to a standard DIMM module.

### 3.4.2.3 Threaded Modules

The threaded module is an alternative to dual-channel modules, and provides a similar efficiency gain without the need for an additional C/A port. As discussed previously, many of today's DRAM components implement a high core-prefetch



ratio. The DDR3 architecture uses a burst length of 8bits transferred at double data rate (on the rising and falling edge of the clock). During the transfer time of a single DRAM access, there is opportunity to issue up to 4 row activate, read, or write requests (using 1T timing and the auto-precharge Since only two commands function). (activate and read or write) are required to access a burst of data, there is excess command bandwidth to issue an access to another memory "thread". For a future DRAM with a core-prefetch ratio of 16, there is sufficient bandwidth to implement two commands even with 2T C/A timings.

A threaded module can implement dual memory channels on a memory module while sharing a single C/A bus. Commands to the two memory channels are multiplexed across the same set of signal traces as a traditional single-channel module but with separate chip selects to differentiate the command to the respective memory channel. The threaded module utilizes the excess command bandwidth and implements a dual-channel module with only the addition of chip-select pins.

### 3.4.3 Threading System Benefits

### 3.4.3.1 Power advantage

Threaded and dual-channel modules can lower the power of main memory accesses. As shown in the timing diagram below, a threaded or dual-channel module can interleave commands to each respective memory channel. In addition to smaller minimum transfer sizes, threading also reduces row activation power compared to a conventional DIMM module.

For a conventional eight-device module, all eight DRAMs are activated (ACT) followed by a read or write (COL) operation on all eight devices. A threaded or dual-channel module can accomplish the same data transfer by activating only four devices and then performing two consecutive read or write operations to those devices. Since only four devices are activated per access instead of 8 devices on a conventional DDR3 module, a threaded or dual-channel module achieves equivalent bandwidth with one-half the device row activation power. On a memory system, this translates to approximately 20% reduction in total module power.





A threaded module was implemented using DDR3 DRAMs. Measurements of this threaded module running an IDD7 equivalent pattern confirm the 20% power savings of the threaded module compared to a conventional single-channel module.

Rambus.

	IDD7 Current (VDD+VDDQ)	Module Power
Threaded Module	2.133 amp	3.20Watt
Conventional Module	2.668 amp	4.00Watt

Table 3.Power Comparison of a<br/>Conventional and<br/>Threaded DDR3 Module

### 3.4.3.2 Bandwidth Improvement from Threading

Another benefit that threaded or dualchannel modules offer is increased sustained bandwidth at high data rates. Many modern industry-standard DRAMs have limited bandwidth due to power restrictions on the DRAM devices. On DRAMs starting with the DDR3 generation, only a limited number of banks may be accessed in order to protect the on–DRAM power delivery network and maintain a stable voltage for the memory core. This parameter, know as tFAW (Four Activate Window period) allows only 4 banks to be activated in the rolling tFAW window.

The bandwidth limitations of tFAW begin to affect DDR3 DRAMs at data rates above 800Mbps. The tFAW parameter relates to the core access time parameter tRC and the fact that DRAM data rates are scaling faster than core access times.

For a computing system, tFAW restricts the memory controller from issuing additional row activate commands once four activates have already been issued in a given tFAW period. This stalls the memory controller and results in lost data bandwidth. A DDR3 DRAM 1066Mbps running at loses approximately 25% of its sustained data bandwidth since memory transactions are stalled by tFAW restrictions. At 1600Mbps data rates (shown in the figure below), a DDR3 DRAM loses up to 50% of its sustained data bandwidth since memory transactions are stalled by both tFAW and



tRRD (RAS to RAS row operation delay) restrictions.

Rambus

Since the DRAMs in a threaded module are activated half as often as those in a conventional module, the sustained bandwidth of a threaded module is not limited by the tFAW or tRRD core parameters.

### 3.4.3.3 Improved Efficiency From Increased Number of Logical Banks

Threaded or dual-channel modules can also increase random bank performance since the architectures increase the number of logical banks versus conventional singlechannel modules. For instance, memory accesses from different threads can be mapped to independent banks which decrease the chance of bank conflicts.

#### 3.4.3.4 Module-level Threading Summary

The Rambus threaded and dual-channel modules improve transfer efficiency and provide smaller access granularity through an improved module topology. These module-level innovations can reduce row access power by half, resulting in a 20% reduction in module power. Threading also improves module bandwidth for 64Byte transfers by up to 50% compared to current DDR3 modules.

### 3.4.4 Micro-threading:Device Level Threading

Elements of the Rambus threading approach at the module level can be applied to the DRAM devices using an innovation called "Micro-threading." Micro-threading improves transfer efficiency and more effectively uses DRAM resources by reducing row and column access granularity. It allows minimum transfer sizes to be four times smaller than typical DRAM devices by using independent row and column circuitry for each quadrant of physical memory space. Independent addressability of each quadrant also complements the threaded memory workloads of modern graphics and multicore processors. This unique architecture enables micro-threading to maintain the total data bandwidth of the device while reducing power consumption per transaction.

### 3.4.4.1 Core Operation

As discussed previously, the larger prefetch ratio of modern DRAMs have increased the access granularity (a measure of the amount of data being accessed). Many DRAM subsystems deliver more data than

	Bytes Transfers		
	Conventional Modules	Threaded Modules	Bandwidth Gains/ Threaded Modules
DDR3-1066	512Bytes	640Bytes	25%
DDR3-1333	512Bytes	640Bytes	25%
DDR3-1600	512Bytes	768Bytes	50%

Table 4. Bandwidth Improvement of a DDR3 Threaded Module

necessary for applications such as graphics or multi-core computing. Retrieving excess data is inefficient and wastes DRAM and signaling power.

Rambus.

Access granularity is a function of the accessibility of data within a memory architecture. A typical DRAM is comprised of eight storage banks. Within such DRAMs, each bank is typically further subdivided into two half banks, "A" and "B". For such a DRAM with 32 data pins, each A half bank is connected to 16 data pins and each B half bank is connected to 16 pins. The banks are in opposite quadrants of the physical die, and each quadrant has its own dedicated row and column circuitry - each bank half operating in parallel in response to the row and column commands.



DRAM Core

A row command selects a single row in each half of the bank being addressed, thereby sensing and latching that row. Physical timing constraints impose a delay (i.e., tRRD) before a row in another bank can be accessed. Column commands are similarly constrained (i.e., tCC). However, the row timing interval is typically twice the column interval; therefore two column timina commands can be issued during the mandatory delay required for a single row activation.

The column prefetch length, the amount of data delivered per transaction, is determined by the respective column and row timing delays and bit transfer rate, where:

Prefetch = timing delay/bit transfer rate

A core of a mainstream DRAM typically operates up to 200MHz, whereas a core of a high-performance industry-standard DRAM can typically operate up to 400MHz. Core frequencies exceeding 400MHz are difficult to achieve using modern industry-standard DRAM technologies without sacrificing production yields or increasing costs. Therefore, a column prefetch of 16bits is required for such a high performance DRAM core to support external data rates exceeding 3200MHz since the DRAM core is organized with each half-bank operating under the same row or column operation.

### In addition:

Column granularity = (column prefetch) x (number of data pins per half bank) x (number of half banks per access)

### Or:

For a 32bit wide DRAM with 16 data pins per half bank:

Column granularity per access =  $16 \times 16 \times 2 = 512$  bits or 64Bytes.

Moreover, during the row timing interval, in order to maintain peak bandwidth, at least two column operations must be performed. This is typically described as two column address strobes per row address strobe (two CAS per RAS). This results in a minimum row granularity of 128Bytes. This large access granularity translates into inefficient data and power utilization for applications such as 3D graphics.

### 3.4.4.2 Micro-threading in DRAM Cores

Using largely the same core resources as in the previous example, a sample microthreaded DRAM core has 16 banks, each bank in the micro-threaded DRAM being equivalent to a half bank in the typical DRAM core. The even numbered banks connect to the A data pins and odd numbered banks connect to the B data pins (again with 16 pins in each case). However, unlike a typical core, each four-bank quadrant can operate independently, through the use of independent row and column circuitry for each quadrant. Moreover, interleaving, simultaneous access to more than one bank of memory, allows concurrent accesses to the lower quadrant on the same physical side of the core as the previous access.



### Figure 31. Micro-threaded 16-Bank DRAM Core

Micro-threading enables four independent accesses to the DRAM core simultaneously. Although the same time interval as a typical core must still elapse before accessing a second row in a particular bank or bank quadrant, the three banks in the other quadrants remain separately accessible during the same period. Columns in rows in quadrants can be concurrently other accessed even though a column timing interval must pass before a second column is accessible in the previously activated row. The net effect of this quadrant independence and interleaving is that four rows (one in a bank of each quadrant) and eight columns (two in each row) are accessed during the row timing interval (compared to a single row and two columns with the typical DRAM technique).

# Rambus.

### 3.4.4.3 Relaxation of DRAM Core Parameters

Timings are similar to the typical DRAM core, but each column sends data only for half the column timing interval. The interleaved column sends data for the other half of the interval. Micro-threading reduces minimum transfer granularity size while maintaining a high-yielding and cost effective core frequency. By interleaving the column accesses from four different banks, a microthreaded DRAM core (of a given column prefetch length and core frequency) can support a data rate two times higher than that of a conventional DRAM core. Conversely, micro-threading of the column operation enables a DRAM core to costeffectively sustain a specific data transfer and granularity while relaxing the column cycle time (tCC) by up to two times compared to those of a conventional DRAM core.

With micro-threading, column granularity is now:

Column prefetch/2 x 16 pins =  $16/2 \times 16 =$  128bits or 16Bytes (one quarter of the

previous value).

The row granularity is 32Bytes (again one quarter of the previous value). Microthreading's finer granularity results in a performance boost in many applications. For example, in a graphics application with 8Byte micro-threaded column access granularity, computational and power efficiency increased from 29 to 67 percent after introducing the technique.

### 3.4.4.4 Increased Number of Logical Banks

Similar to threaded or dual-channel modules, micro-threading can also increase random bank performance of a DRAM by increasing the number of logical banks versus conventional non-threaded DRAMs. For instance, memory accesses from different threads can be mapped to independent banks, making use of double the number of logical banks and thus decreasing the chance of bank conflicts.

### 3.4.4.5 Micro-threading Summary

The Rambus micro-threading innovation improves transfer efficiency, smaller access granularity, with largely the same DRAM







core. Micro-threading maintains high sustained data bandwidth with transfer sizes at one-half or one-quarter of those of conventional DRAMs while lowering power consumption. For small accesses, microthreading enables twice the data rate from a DRAM core over conventional techniques.

Micro-threading offers a scalable path for future data bandwidth increases and higher core-prefetch ratios. Micro-threading is applicable to industry DRAMs and is incorporated in the XDR2 memory architecture.

### 4 Summary

Advances in multi-core computing, virtualization, and converged graphics/ general-purpose processors are increasing the requirements for future memory systems. This whitepaper presented an architectural concept for a future memory system that addresses the anticipated demand for higher bandwidth with higher data rates, lower power, higher throughput and higher capacity.

The architecture incorporates Rambus FlexClocking technologies to support data rates of 3200Mbps with very low standby power and fast turn-on time. The clocking architecture is strobeless and employs Rambus FlexPhase timing adjustment circuits to transmit and receive data. The memory controller can be implemented with optional compatibility to DDR3 memory. Beyond reducing standby clocking power, the architecture also incorporate Near Ground Signaling to reduce IO power by up to 80%, and dual-channel or threaded modules to reduce row access power by 50%.

Threading technologies also improve bandwidth and efficiency without the need to increase the processor cache line size. Dual-channel modules, threaded modules or micro-threading technologies can support 64Byte memory transfers at full bus utilization, resulting in efficiency gains of up to 50% compared to current DDR3 memory systems.

To address higher capacities, the architecture incorporates Rambus DPP technology, which can double or quadruple capacity that is natively supported by a memory channel while maintaining the ideal point-to-point signaling topology for high data rates. DPP also enables module upgradeability without the added power, cost, and latencies associated with buffered approaches.

The future memory architectural concept, employing Rambus innovations, doubles data rates over current DDR3 memory and substantially improves throughput and transfer efficiency with 64Byte accesses. The architecture reduces standby, IO and core power, and can natively support higher capacities per memory channel.

### Authors:

Tom Giovannini Scott Best Ian Shaeffer Kendra De Berti Michael Ching

Rambus Inc.

NOTICE © Copyright 2009 Rambus Inc. 4440 El Camino Real Los Altos, CA 94022 U.S.A. All rights reserved.

Copyright in this document is owned by Rambus Inc. The use of this documentation is governed by an agreement containing restrictions on use, access, and disclosure. Rambus Inc. reserves the right to make changes to this documentation without obligation to notify any person or organization. No part of this document may be photocopied, reproduced, transmitted, transcribed, stored in a retrieval system or translated to another language, in any form or by any means, electronic, mechanical, magnetic, optical or otherwise, or disclosed to third parties without the prior written consent of Rambus Inc. Rambus has issued and/or pending patent applications relevant to technologies discussed in this document.

Rambus, RaSer, FlexPhase, SO-RIMM, RIMM, RDRAM, and the Rambus Logo: () are trademarks or registered trademarks of Rambus Inc. in the United States and other countries. Rambus and other parties may also have trademark rights in other terms used herein.

THIS DOCUMENT IS PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, OR NONINFRINGEMENT. THIS DOCUMENT COULD INCLUDE TECHNICAL INACCURACIES OR TYPOGRAPHICAL ERRORS. CHANGES ARE PERIODICALLY ADDED TO THE INFORMATION HEREIN; THESE CHANGES WILL BE INCORPORATED IN NEW EDITIONS OF THE DOCUMENT. RAMBUS INC. MAY MAKE IMPROVEMENTS AND/OR CHANGES IN THE TECHNOLOGY DESCRIBED IN THIS DOCUMENT AT ANY TIME. NO EXPRESS OR IMPLIED LICENSE IS HEREBY GRANTED UNDER ANY PATENTS, COPYRIGHTS, MASK WORK RIGHTS, TRADEMARKS OR OTHER PROPRIETARY RIGHTS.