EXPANSION I/O

The right platform choice for your current, legacy and future application needs

Improve TCO with flexibility for scale computing and I/O. Why Magma solutions and Rack Disaggregation (RD) can save money, insulate integrators from change and offer better performance for many applications.

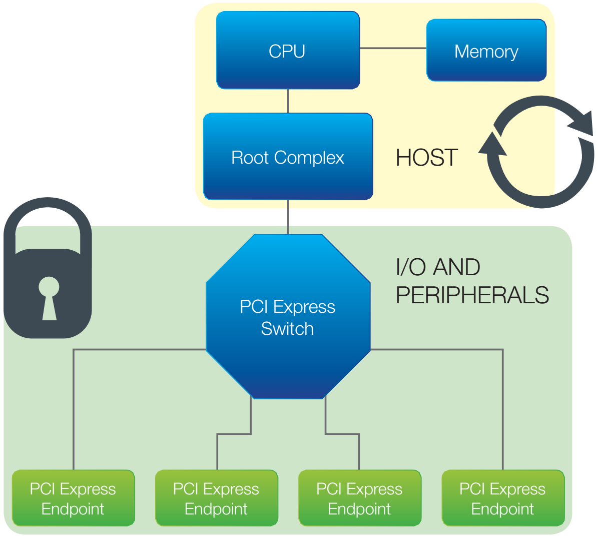
MANAGING CHANGE

Platform decisions are critical for a system designer for a number of reasons. Initial cost of a system is a key factor, but for professional applications that often have product life cycles measured in years and even decades, the choice of the underlying architecture has a much larger implication in operational expense, and flexibility in the solution to support change where needed, and maintain stability in other elements of the solution. Depending upon the application, a system designer may desire for processing performance to improve over time, with I/O resources remaining constant, or the inverse may be true, where processing needs to remain consistent over time, with improvements in I/O the area where innovation and improvement need to be accommodated. In this paper, we will explore the flexibility best offered by a disaggregated I/O model (I/O separated from processing), the drivers of change, and the promise of new capabilities being enabled by innovations in the storage and GPU markets.

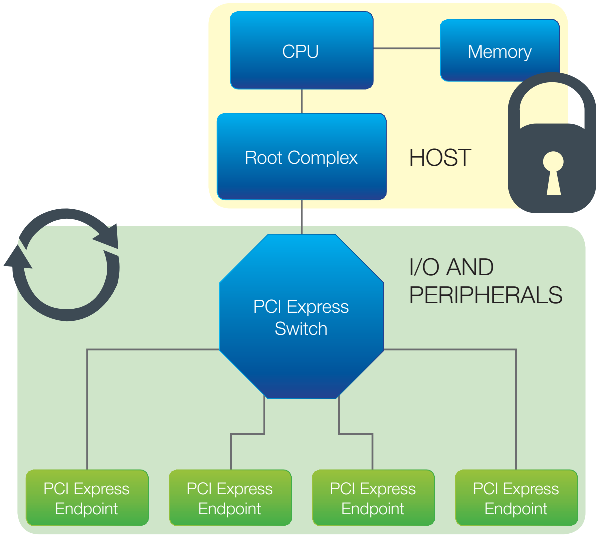
The design for today's PC and server is evolving from a traditional all-in-one motherboard design to a disaggregated I/O model where peripheral cards are physically located in an expansion chassis, and the processor and memory subsystems are contained in the host platform. In the PC domain, the best example of this is what Apple is doing with the Mac Pro: using Thunderbolt to provide access to remote peripherals.

In server architectures, the term Rack Disaggregation is used to capture the concept of a root processing and memory ‘server’ attached to a separate I/O expansion chassis through, most typically, PCI Express (PCIe). Expansion solutions such as those provided by Magma allow flexibility for designers and integrators not only how an original system is architected, but also in how that system evolves with generational changes in the underlying technologies, such as processing, memory, storage and I/O throughput.

It is equally important for many applications to NOT have to change for arbitrary reasons such as a vendor obsoleting a motherboard or I/O peripheral. By using expansion I/O, a system integrator can choose to evolve parts of the system while keeping other elements fixed architecturally. In many cases, specialized cards are designed around a certain technology (e.g. classic PCI), and the system integrator does not have the time, money, or desire to convert an older design into a newer technology (PCI to PCIe for instance), just because a motherboard vendor now no longer supports PCI slots. As an example, we have some customers that have designed their own specialized I/O peripherals costing $30K or more per card. By continuing to offer classic PCI expansion solutions today, we help insulate our customers from undue change for cards they might have designed 15 years ago, yet allow them to migrate their host processing to contemporary state of the art solutions. With Disaggregation, the integrator has the choice to keep certain portions of the overall solution fixed, and/or upgrade different portions at different rates:



When a fixed I/O subsystem is critical and the ability to modify the host is desirable, disaggregation provides this flexibility.

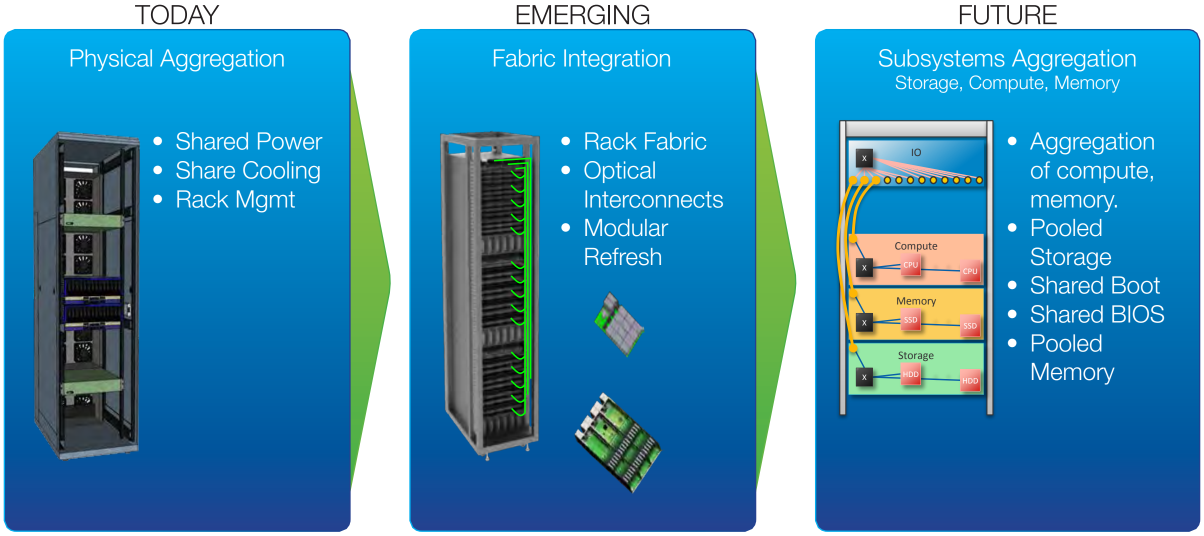


When a fixed host is critical and the ability to modify the I/O subsystem is desirable, disaggregation provides this flexibility.

SERVER RACK DISAGGREGATION (RD)

The larger server vendors, such as IBM, Dell and HP, as well as silicon providers such as Intel, are driving the concept of Rack Disaggregation (RD) within the server/data center market. This is due to the need for flexibility, serviceability and performance scalability of the server, allowing processing elements, memory, I/O and storage to evolve each at their own pace. Silicon advances are well known to follow “Moore’s Law” (doubling of silicon density roughly every 18 months), but I/O channel and storage have their own pace of advancement that are slightly different. Practical market issues such as vendor obsolescence of peripheral card solutions also conspire to make a system integrators roadmap support difficult to manage over the balance of time. By disaggregating I/O from the core CPU, each of these elements can migrate over time without having to do wholesale forklift upgrades of the complete system.

Evolution of Rack Disaggregation


TODAY

Physical Aggregation

- Shared Power
- Share Cooling
- Rack Mgmt

EMERGING

Fabric Integration

- Rack Fabric
- Optical Interconnects
- Modular Refresh

FUTURE

Subsystems Aggregation

Storage, Compute, Memory

- Aggregation of compute, memory.
- Pooled Storage
- Shared Boot
- Shared BIOS
- Pooled Memory

Platform Flexibility > Higher Density > Higher Utilization

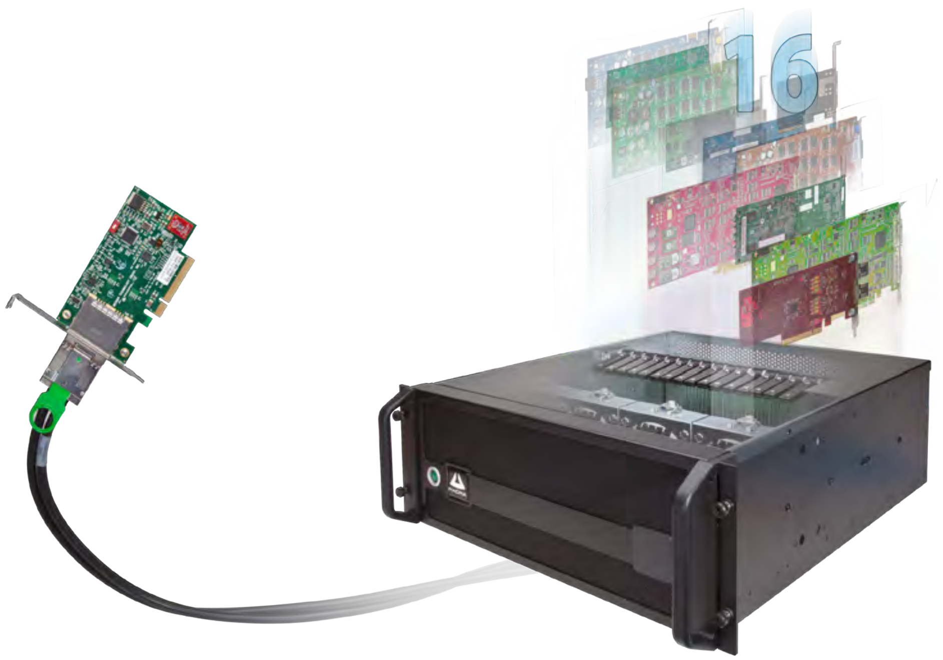
Source: Intel

PRACTICAL MOTHERBOARDS

What this means is that server motherboard designers are tending to optimize the motherboard for processor and local memory utilization and relying on expansion connection for I/O card support, including GPU and mass storage. In some cases, a single, half height PCIe slot is all that remains for expansion, with the expectation that this will be for a PCIe-PCIe link interface to an expansion chassis.

”Traditional” Server Motherboard with multiple PCI and PCIe slots

Contemporary Server Motherboard with 1-2 PCIe Expansion slots



Magma EB16 PCIe Expansions Chassis shown with link cable and interface for server

PERFORMANCE IMPROVEMENT

Performance can also be improved by the hierarchical nature of a disaggregated I/O model. PCIe switching of PCIe-PCI bridging provides an isolated logical layer for inter- peripheral communication. For certain applications, this is a distinct advantage in that any activity on the root PCIe layer on the server will not mix with inter-peripheral communication on the expansion chassis and vice versa. A real time OS running on a host server might be sensitive to this activity in terms of latency and/or the activity on the lower layer PCIe hierarchy might be sensitive to activity on the host. By providing a local switch on our motherboards, we allow for this hierarchical isolation of local bandwidth. We'll look later at how new classes of computing with clustered GPUs are best offered as an expansion solution.

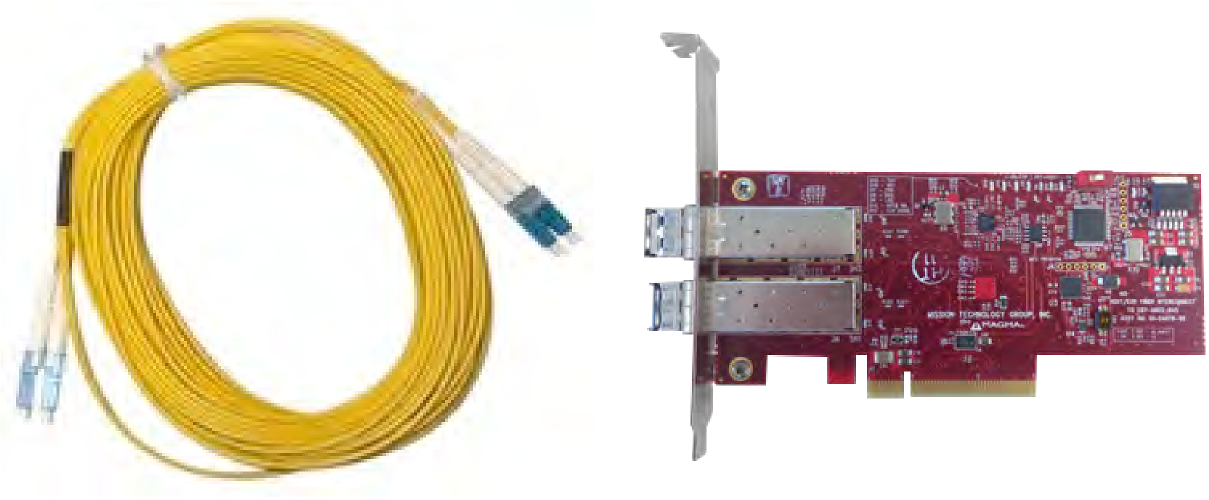
REALLY DISTANT (AND QUIET) I/O

A compelling new capability that disaggregated I/O provides is the utilization of fiber link layers between the host and the I/O subsystem. This has a couple of key benefits:

1. Long distances between the server complex and the I/O subsystem (up to 300 meters). We have some applications where the server is the stern of a ship, and the I/O is forward in the bow on an aircraft carrier.

2. Fiber is electrically quiet in terms of emitted EMI or susceptibility to injected EMI by outside sources. This is important for long distances [as in 1) above], but also for shorter chassis-to-chassis or shelf-to-shelf communication links within an electrically noisy data center or central office or in an environment highly sensitive to EMI.

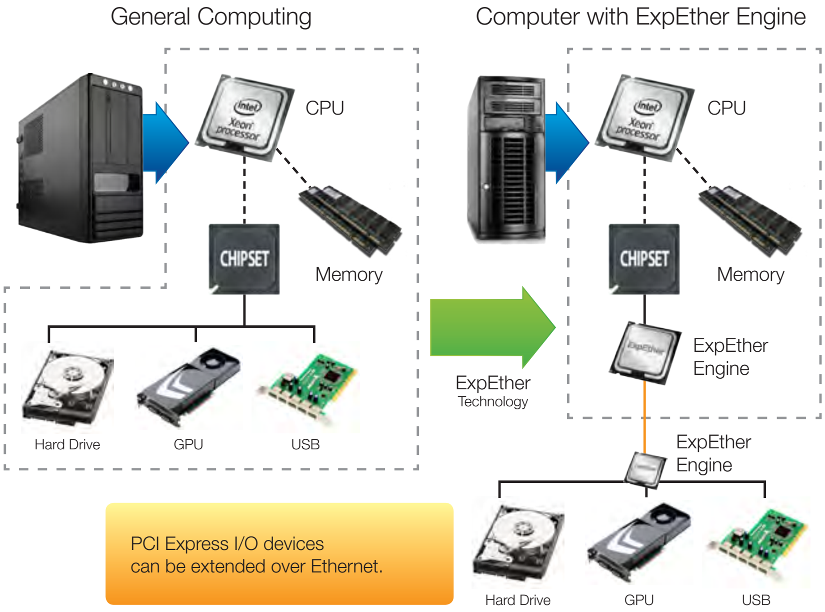
3. In human harsh environments, such as high temperature or loud environments, the server subsystem can be located in a compatible climate controlled area, with the I/O subsystem located more remotely or in tough to access areas.



PCIe fiber link for remote expansion

PCIe OVER ETHERNET

Another new form of PCIe extension is by using Ethernet as a transport mechanism for carrying PCIe, allowing virtualism of the PCI environment through traditional Ethernet switching topologies allowing much longer distance disaggregation.



PCI Express I/O devices
can be extended over Ethernet.

THE DESKTOP AND THUNDERBOLT

Intel and Apple have codeveloped a new cabled expansion solution for remote I/O called Thunderbolt (10 Gbps per channel/20Gps total) and now Thunderbolt 2 (20 Gbps per channel/40Gps total) shipped first on Apple MacBook Pro and will be the I/O mechanism of choice for the Mac Pro.

Apple Mac Pro - Fall 2013

Apple's newer solutions such as the Mac Pro and MacBook products are leading the wave of high end personal computing platforms moving to a disaggregated I/O model via Thunderbolt and Thunderbolt 2. Professional developers involved with image processing, media processing and other applications benefitting from coprocessing elements such as nVidia or AMD Graphics Processing Units or large amounts of fast SSD storage, will now use solutions such as our EB3T and EB1T expansion platforms through a Thunderbolt cable connection.

Magma Expansion Solutions using Thunderbolt

ELEMENTS OF CHANGE

All of these real system and silicon changes are fundamentally changing the way that servers are being built for the data center and PCs used on the desktop. This architectural shift is having a similar impact and benefit on the way that OEM providers and system integrators build their embedded systems. Magma is committed to providing OEM solutions to our customers both in the leading edge of a technology adoption (e.g. PCI Express x16 Gen 3 or Thunderbolt 2, today) as well as for the long term (we still ship classic PCI expansion solutions). Investment is preserved for solutions based on one technology (e.g. PCI) needing to be supported by newer platforms that no longer support an older technology. In fact, with our solutions, it is possible, assuming drivers exist, to communicate between a brand new Thunderbolt 2 enabled system and an old classic PCI card designed in 1994.

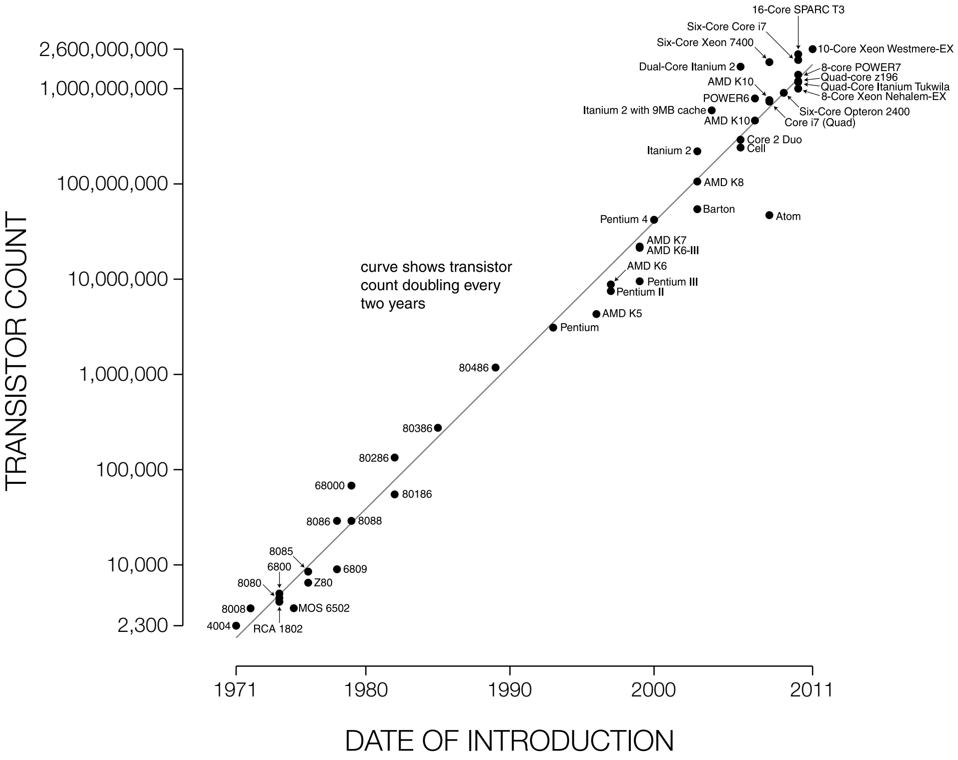
Over the balance of time, several factors conspire to inject change into the life cycle of a released product. Silicon, storage and I/O interconnect improvements generally create faster and cheaper solutions on a logarithmic curve. Each technology has a slightly different pace however, and the changes tend to come in chunks...Intel will improve a process from 22nm to 18nm, not one nm at a time, and correspondingly, actual processor chips come out at discrete intervals. The same is true for storage and I/O interconnect. As each evolves, specific vendors make choices to create released products and generally replace older products with newer and improved products (e.g. a new motherboard, or a new high density storage disc). For a system integrator, as long as it is functionally equivalent, this is fine, but the reality is that traditional server motherboard where all of the elements are combined, change happens across the board (literally!) and I/O slot types change when processors change. Disaggregation of I/O does two things: it creates a distinct I/O subsystem and it creates a distinct link layer that can evolve as well.

Progress in technology is generally viewed and valued as a good thing. Faster, smaller, lower power, increased density, lower cost systems from generation to generation have been driving the computing industry for decades. For any given project or product, change is not necessarily desired once the engineering development effort has been done and the product is in the market. Practical realities of vendors obsoleting products or elements of products, create unwanted rework of portions of released products, often at inopportune times and at the expense of other value added work. Disaggregating I/O allows for more managed change (instead of wholesale rework) and a longer life cycle for products, which saves money over time.

SILICON ADVANCES (MOORE’S LAW)

MICROPROCESSOR TRANSISTOR COUNTS

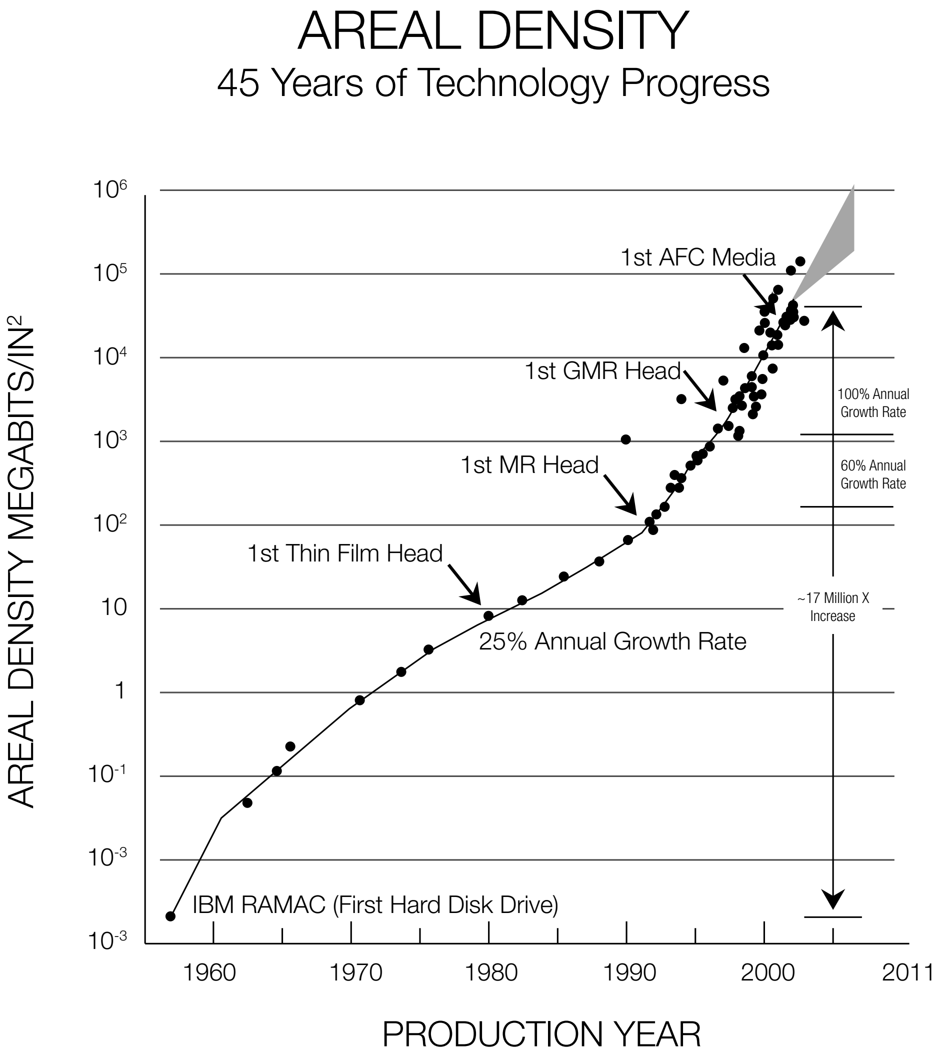
1971–2011 & Moore’s Law



Source: Wikipedia Commons

STORAGE ADVANCES

Areal Density for storage follows a similar trend of logarithmic growth over time, but is different by a degree to Moore's law. While this might seem not to be a big distinction, the practical change of storage density being different than processing improvement causes churn for systems integrators because while graphs look nice and tidy, actual product advances tend to come in chunks.



Source: Tom's Hardware

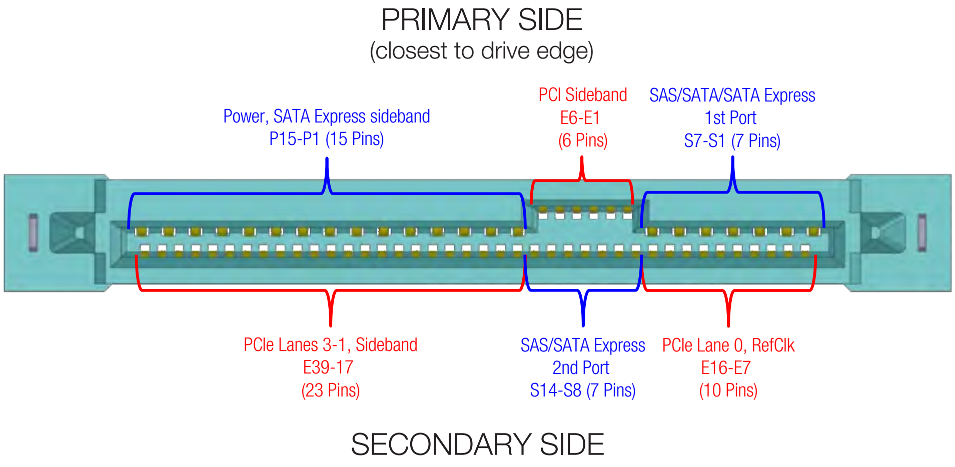
THE STORAGE SHIFT TO SOLID STATE DISK (SSD) SOLUTIONS

Within the storage market, there are a number of fundamental shifts occurring that benefit and/or are benefitted by the general move to disaggregation:

1. Rotating media being replaced by faster Solid State Disk (SSD) technology. The performance advantages outweigh the higher cost (for now) per bit and the ruggedness of SSD solutions is also of benefit to a variety of applications.
2. The move to direct connect PCIe interfaces, removing the SATA/SAS controller requirement, moving the storage logically closer to the processor. This removes latency and faster overall system throughput by removing potential bottlenecks.

FusionIO 1.6 TBybte Direct PCIe Attached SSD

Creation of a new connector (SFF-8639) for 1.8/2.5/3.5" SSD drives, including support for dual porting (a key feature of SAS) for high availability applications.



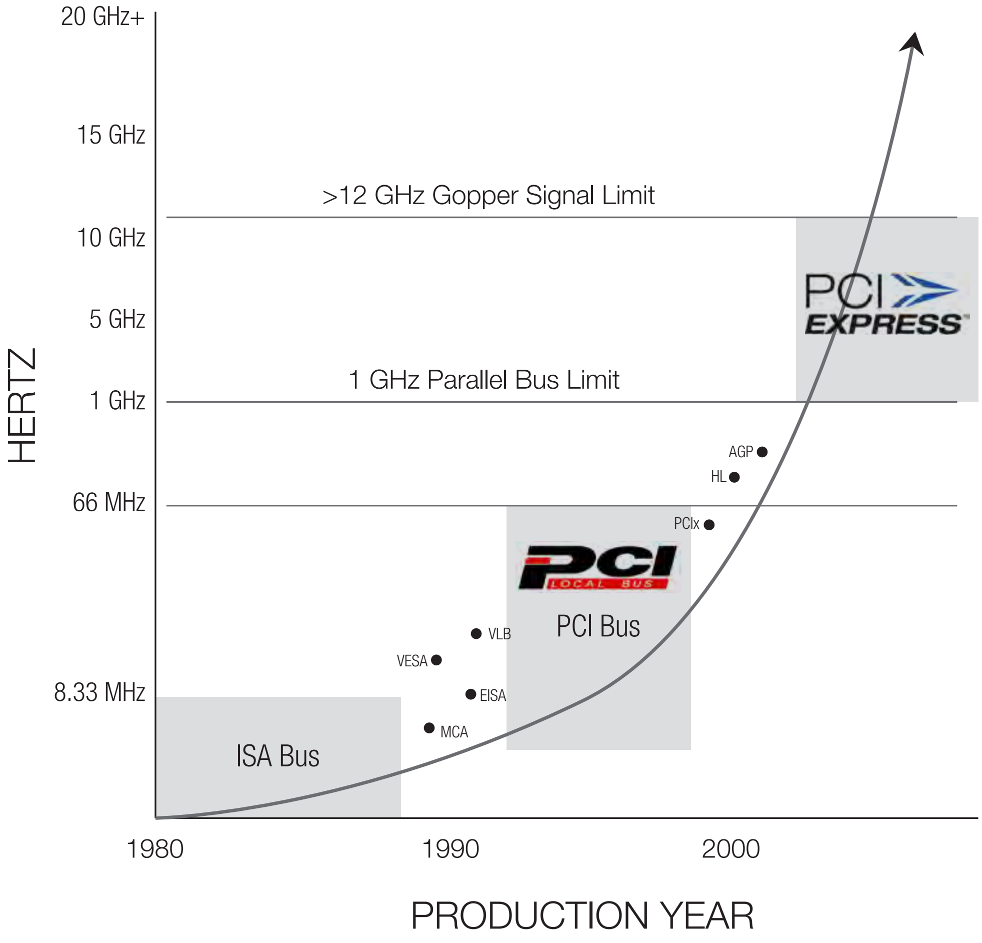
SFF-8639 Connector

Samsung 1.6 TByte SSD

These changes in the storage market are game changers in a lot of ways: speed, density, removal of complexity (SAS/SATA control layer being replaced by NVM Express). The shift to adoption of direct PCIe connection, all allow for some fascinating new topologies to be constructed (and the subject of another white paper). The key relevance to our customers is that we will be able to offer these new technologies alongside our traditional PCIe solutions as well as legacy I/O such as classic PCI or PCI-X.

I/O ROADMAP

Similarly, I/O link speeds for things like DRAM access, or chipset interconnect follow an advancement curve, but tend to lag silicon and areal density advances.



At Magma, we are building and providing solutions coincident with silicon availability for PCIe Gen 1/2/3 and Thunderbolt 1/2, and we will continue to offer solutions in lockstep with the market as Gen 4 rolls out.

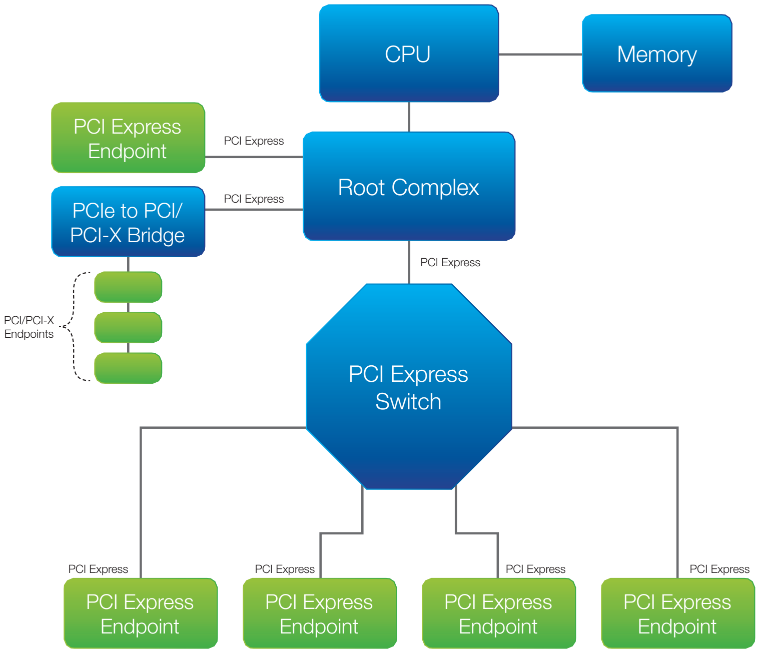
OBSOLESCENCE OF SYSTEM ELEMENTS

Many of today's system configurations are composed of peripheral cards and servers that have a relatively short lifetime in the market. For professional OEM applications, disaggregation of the I/O complex (Magma Expansion Chassis and peripheral cards within) from the compute platform (Server) allows each to evolve at different paces. A server motherboard vendor may offer different combinations of PCI and PCIe slots with each successive generation of their product line, including not offering older PCI slots at all on newer generation motherboards. While the cost of having to retrofit older configurations with new solutions is an engineering cost relatively easily identified, the true hidden cost that is harder to enumerate is that these events (obsolescence) are often a surprise and difficult to support at the expense of other projects and often at a non opportune time. This is opportunity cost, which means that effort spent on redesigning older solutions just because a motherboard can no longer be purchased with a PCI slot, for example, comes at the expense of developing some other project that truly adds value.

When peripherals are forced to be redesigned, additional effort is often spent in refactoring device drivers and retesting of system configurations. By retaining the ability to use older I/O peripherals with newer host processors, much of this forced change can be avoided.

PERFORMANCE

We reviewed earlier how PCIe, through the inherent isolation provided by PCIe switching elements, can offer overall higher system performance by allowing interprocess communication on one layer (e.g. direct DMA between peripherals) to be less impeded by traffic on other layers.

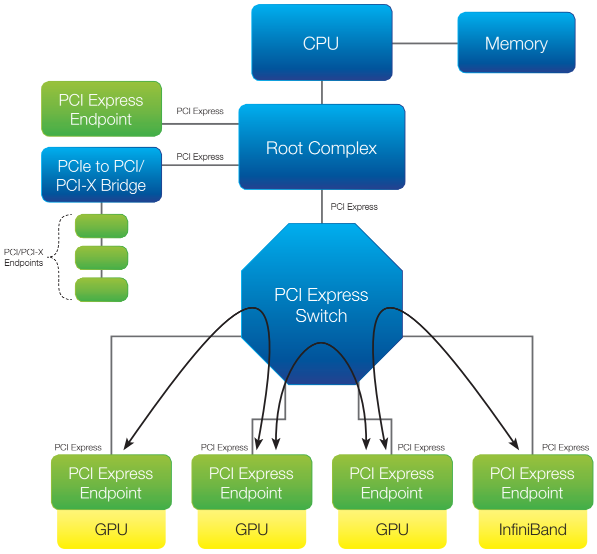


PCI Express Switch Hierarchy

Fast Gen 3 connections between endpoints and optimized arbitration schemes within today's contemporary switch silicon can offer better performance than if the endpoints are directly connected to the root complex, which often has other housekeeping (e.g. LAN source/sinking) tasks to attend to. By moving this traffic below the switch in the hierarchy, overall system bandwidth is improved by segregating traffic into homogeneous domains for better interprocess communication.

WHOLE NEW CATEGORIES OF PRODUCTS
GPUS / HIGH PERFORMANCE COMPUTING

An exciting new product category being enabled by the incredible parallel processing capability of Graphics Processing Units (GPUs) is the ability to cluster GPUs together with the super fast interconnect provided by PCI Express. The GPU market is fueled primarily by the graphic rendering and gaming markets, but the side benefit of the availability of these products for a reasonable price is that they can be dedicated for other types of applications needing repetitive array type processing, such as simulation applications or genome processing.

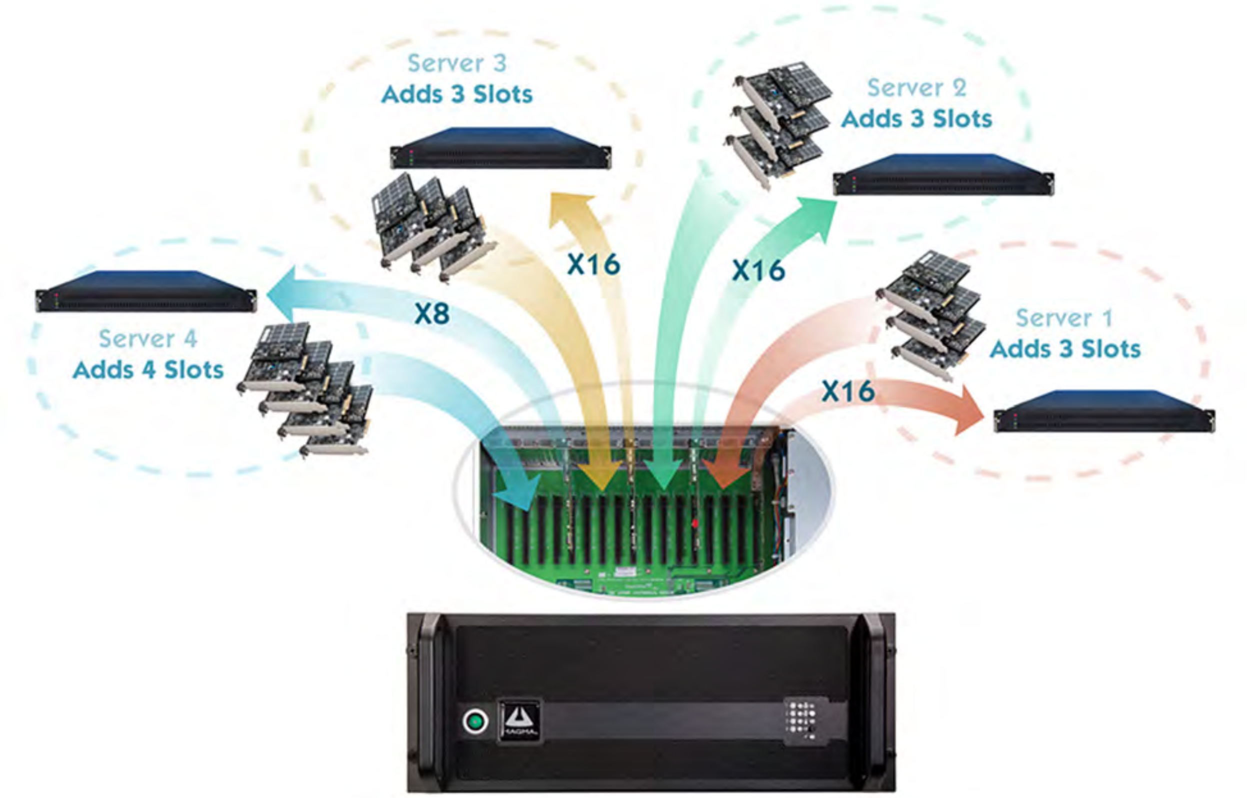


Direct Memory Access and RDMA (InfiniBand) communication

Supercomputer level computing has reached the desktop, and today's solutions will be dwarfed by tomorrows capabilities given the dynamics of silicon, storage and I/O change discussed above. Using a disaggregated I/O solution allows these clusters of GPUs to upgrade over time, including future generations of PCIe (Gen 4) and Thunderbolt.

PARTITIONING

Within PCI Express switching, there is a capability to partition the PCI Express topology such that multiple hosts can communicate with peripherals through the same switch silicon but be fenced off (isolated) from other segments. This allows an expansion chassis to have multiple hosts connected to it, each with a local private set of peripherals.



Partitioned PCI Express Chassis

Such a system allows an integrator to save rack space by combining I/O peripherals into one chassis that shares common power, cooling and physical space.

CONCLUSIONS

Rack disaggregation is about flexibility and control over your platform requirements and the manageable evolution of your platform over the lifetime of your application. Typical embedded and professional server - I/O implementations will need to be supported for years and even decades, in some instances, as compared to the sub-year pace of the traditional PC or server market. Magma's solutions leverage the general trend within the desktop and server markets and we offer leading edge technologies (today: PCIe Gen 3 and Thunderbolt and Thunderbolt 2) coincident with their availability in silicon. By supporting older expansion chassis's such as classic PCI with PCIe based hosts or even Thunderbolt, we provide platform stability for system integrators, protecting their investment in peripheral card development for specialized applications.

The inherent isolation of PCIe switching technologies can be used to advantage for overall faster system performance. Super Fast and low latency interconnect between peripheral devices such as GPUs are creating whole new categories of system level products, bring supercomputing level performance down to the desktop.

Technology elements evolve at different rates and tend to be chunky as to when they become available, By separating the host processing element and peripheral I/O, each can evolve more naturally. Obsolescence by vendors is mitigated because each functional element (processor, link, I/O subsystem) tends to be replaced by a functional equivalent that is faster or has more capacity, but also backwards compatible. The older way of building a motherboard with processor, memory, I/O and storage integrated together tends to churn one element when the others change (e.g it is hard to find a multicore motherboard that supports a lot of classic PCI slots today). Disaggregation divides the system topology into manageable subsystems.

Flexibility in mixing Thunderbolt 1, Thunderbolt 2, Gen 1, Gen 2 and Gen 3 PCIe hosts and I/O expansion can allow a Gen 3 PCIe based server to talk to an older Gen 1 PCIe I/O expansion chassis or an even older PCI based expansion chassis. The converse is possible. An old Gen 1 based server can interoperate with a fast Gen 3 based expansion chassis, where the I/O cards are all communicating x16 Gen 3 to each other, yet the old host might only have a Gen 1 x4 link. There are practical challenges in driver support and BIOS capabilities for specific vendors, but the architecture of our solutions allow for a myriad of solutions and we support them for much longer than the traditional PC or server vendor does.

Disaggregation of I/O also allows for distance separating of the I/O subsystem from the server, which may be a requirement for installations where the data center for computing is on one area (e.g the stern of a ship) and the I/O peripherals are far remote and reach via a fiber link between chassis or maybe a PCIe over Ethernet routing.

All of these factors help manage change and wholesale forklift upgrade requirements of a platform. Incremental change is much easier to manage and much less costly of the life of a product that can measures 5-10+ years for many embedded applications.

Our longtime expertise with PCI, PCI-X, PCIe and now thunderbolt technologies position us well to support your long term needs. The exciting changes underway within the storage and GPU technology markets are providing for whole new classes of applications and processing performance. Our solutions help partition a solution differently and offer better longterm manageability of change and easier integration of new technologies as they are available for system designers and integrators. Our focus on high speed interconnect, best represented by PCI Express today, assures our platform users that we will offer best-in-class I/O performance over the balance of time.