HP Moonshot: An Accelerator for Hyperscale Workloads

Sponsored by HP, see HP Moonshot for more information – www.hp.com/go/moonshot

Executive Summary

Hyperscale data center customers have specialized workloads, and they are asking for specialized architectures and equipment to accelerate those workloads, with the goal of increasing data center density while decreasing power consumption and other costs.

In 2010, HP gave its Moonshot team a charter to break out of HP's mainstream enterprise value propositions. Creating differentiation in a commodity market is difficult. Likewise, it is easier to be different in an emerging technology market, mostly because in those early markets, competitive benchmarks are an art, not a science. The hardest part though, is creating sustainable differentiation. HP's Moonshot 1500 System hardware platform is as innovative as we have seen, both in its throughput-oriented architecture and in HP's decision to develop a third-party ecosystem.

HP plans on enabling a variety of partner silicon and component vendors to accelerate hyperscale workloads for customers. This includes the lowest power CPUs and adds to it APUs, GPUs, DSPs, and FPGAs, and at scales those vendors would not be able to access on their own. HP's customers will benefit via broader access to innovative accelerators at a faster pace than HP could achieve on its own. HP's success in bootstrapping and sustaining their Pathfinder Innovation Ecosystem will determine their future in the hyperscale infrastructure market.

Megatrends and Innovations

The data center industry is in the early stages of a profound shift that will be more or less complete by the end of this decade. As with most profound shifts, there is a simple dynamic behind it – clients with increasingly ubiquitous broadband access, wired and wireless, enables people to interact with remote, cloud-based software. Cloud-based software is evolving to learn from the aggregate of people using it – referred to as "Big Data" – to provide much better context than a purely local application. Better context translates into better value, with the additional benefit that those increasingly context-rich services can be available on every device subscribers carry or access.

We believe there are three important new categories of context-rich services – mobile, social, and machine-to-machine (M2M). Service providers are creating new workloads to deploy their new context-rich services, and are building massive new data center capacity to host these new workloads. We call data centers operating at this scale "hyperscale" data centers.

Operations and Infrastructure

Enterprise IT has traditionally focused internally, on reducing the friction of doing business, and therefore is primarily engaged in continuity and cost containment. Hyperscale data centers are profit centers. Their goal is to grow their services profitably, and they continually seek new ways to optimize their compute density – a measure of how much service they can deploy for a given volume of data center – against their rising capital and operational expenses to support their services.

Unlike highly virtualized enterprise IT runtime environments, hyperscale services run individual, specialized workloads at such scale that they do not share infrastructure with other workloads at runtime. Instead of optimizing infrastructure to run any workload at a "least common denominator" of service, hyperscale customers are asking their suppliers for infrastructure that they can optimize for high value and specialized workload classes.

There is a solid return for investing in finding an optimal balance of **density**, **costs**, and **expenses** for each workload class. Given the rapid rate of workload and application evolution, finding optimal performance points will be a continuous process for at least the next few years; it demands flexible hardware and software infrastructure.

Power consumption is a top operational constraint at hyperscale, and not only from an operational cost perspective – supply reliability is critical, ecological impact is growing as a social and workforce consideration, and, at a baseline, thermal management solutions add to cost, complexity (which impacts reliability), and power consumption.

Density and complexity interact in new and interesting ways in a hyperscale data center. Processors first increased density by increasing their operating frequencies – each new process technology yielded faster processors. But in the 1990s that became untenable, and multi-core processors then evolved to provide more compute power in the same package, but that route has also hit a point of diminishing returns. Both routes lead to increasing thermal density – i.e. the processor becomes a hot spot that must be thermally managed, and that adds cost in design, parts, and operations.

New architectural approaches seek to use low power, lower speed processors from mobile markets. To increase compute density, more of these smaller processors are required in a given volume of data center. System designs pack these low power processors carefully to more evenly distribute power dissipation so that less aggressive thermal management solutions can be used, while at the same time they must provide adequate system-level data throughput to support increasing compute density.

Modular components are still desired – modularity aides both serviceability and configuration flexibility, which lowers costs – but modularity must be weighed against optimizing resource balances for specific workloads…typically a zero-sum game.

Services, Workloads, and Applications

Definitions for key terms for this discussion and how they relate to each other:

- *Service*: a set of hardware, software, and communications infrastructure that performs a valued function for a customer.
- *Workload*: a subset of a service that performs a well-partitioned function which can be standardized and generically optimized for a community of hardware and software developers.
- *Application*: in the data center world this is software that conforms to a workload category – a deliverable package of code that is a unique instance of a workload and might call for specialized hardware and software infrastructure optimizations for a given vendor or distribution.

Mobile, social, and M2M are services categories – they define valued functions. These three are the primary high growth opportunities for hyperscale data centers. They will drive the majority of hyperscale server sales for at least the next five years.

Cloud and Big Data are workloads – they are components of services, pieces of infrastructure that contribute to the value of services. Here is a short list of some of the current high volume hyperscale workload categories:

- *Web Front End* – serves HTML pages and simple runtime scripts.
- *Web App Serving* – runs an application on a server instead of locally on an end-point. Most consumers are unaware that many of the apps they have installed locally on their mobile devices are simply front-ends to web apps.
- *Streaming* – serves a continuous stream of data, typically media. Streams can originate in databases or can be generated in real-time.
- *Analytics (Big Data)* – management, analysis, and summary of increasingly larger amounts of incoming data in real-time. This includes Hadoop and other frameworks for managing large-scale system resources in parallel.
- *Cloud* – utility computing services that enable other workloads to scale capacity based on demand. Consumers do not directly use cloud utility services, they use cloud-based products, such as storage and social media.
- *Database* – data storage and retrieval, at scales orders of magnitude larger than most enterprise IT databases can handle.
- *Caching* – the distributed nature of the Internet and hyperscale data centers requires intermediate storage between databases and other workloads.

MySQL, Apache Cassandra, and NuoDB are examples of applications. They are all databases, but they are aimed at different workload needs: MySQL is a relational database, Cassandra is non-relational, and NuoDB is based on SQL but scales like Cassandra – a NewSQL database.

Workload acceleration is an artifact of both system architecture and component-level performance, so I will start at the top.

HP's Moonshot 1500 System

The HP Moonshot 1500 System chassis is similar to a blade chassis, but on steroids. It is a 4.3U (7.5 inches tall) chassis that hosts 45 independent hot-plug ProLiant Servers, all attached to multiple fabrics. Like a blade chassis, the HP Moonshot 1500 System chassis sports shared power, cooling, and management resources for those server cartridges.

Unlike blade chassis, it does not have a single, fixed, and shared interconnect backplane. It contains three independent network fabrics – an Ethernet switch fabric, a storage fabric, and a cluster fabric. Each ProLiant Server has access to all three fabrics.

In the HP Moonshot 1500 System, network access for its server cartridges is implemented as two removable Ethernet switch modules that can be configured for redundancy or maximized bandwidth. The initial switch modules implement 1Gbps links, and each server cartridge may have up to four 1GbE links to each switch, for eight links total.

The cluster fabric is not found in blades, and is an independent local interconnect topology in the shape of a 2D torus – groups of three server cartridges are connected north-south in independent rings and groups of 15 server cartridges are connected east-west in independent rings. There are four high-speed hardware lanes in each direction, for another 16 lanes of bandwidth. It is up to server cartridge vendors to specify lane protocol (i.e. PCI-E, Ethernet, SAS, etc.), which is fascinatingly flexible but calls for careful planning when mixing server cartridges from multiple vendors.

Each server cartridge also has access to four SAS or SATA storage lanes, two of which are routed to a central storage resource and two are routed to other server cartridges in a topology that deserves more attention than I can give here. HP designed-in modular disk sharing and the ability to share slices of drives across this independent storage fabric.

The result is that HP separated global and rack-level network traffic from local storage traffic and server cartridge-to-cartridge traffic, providing an opportunity to substantially increase system throughput.

HP ProLiant Moonshot Server Cartridge and Market Enablement

HP's first server cartridge for the HP Moonshot 1500 System will be based on Intel's Atom S1260 processor and is aimed at workloads within dedicated hosting services. The Atom S1260 has two integrated 1GbE ports; each one is routed to a separate Ethernet switch. Because the initial target hosting services do not need additional network or storage bandwidth, this first server cartridge does not use those independent interconnect fabrics. Thread sharing is also not an issue for the workloads enabled by this first server cartridge. HP outfitted each server cartridge with one processor, 8GB of RAM, and from 200GB SSD to 1TB HDD dedicated storage. Typical power for a

completely populated system will be in the ~850W ballpark. That system powers 180 x 2.0GHz threads, with 2GB of RAM for each thread, at under 5W per thread.

Customers can optimize their ProLiant Moonshot server cartridges for hosting applications by adjusting local mass storage capacity on each module and by experimenting with chassis-level network switch configurations, for instance to focus on web front end performance. It can also be configured as an integrated solution for a Web store in a box. This first server cartridge looks well targeted to its audience.

Later in 2013, HP will ship server cartridges with processors from several other processor vendors – AMD, Applied Micro, Calxeda, Intel and Texas Instruments. Each new server cartridge will target specific services, such as high performance computing (HPC), gaming, telecommunications, finance, and genomics research, and workloads like memory caching, web serving and acceleration, Big Data analytics, facial recognition, and video analysis.

For many services it will make sense to mix different types of server cartridges within an HP Moonshot 1500 System to reach an optimal mix of compute, storage, and networking for component workloads. HP intends to assemble a catalog of HP ProLiant Servers, powered with chips from different manufacturers, with widely varying performance and features, and with overlapping new introductions and update cycles. That will be new for our industry. These server cartridges will implement accelerators using a variety of technologies: x86 and ARM CPU cores, compute offload via GPU, DSP, FPGA, and fixed function logic, and eventually APUs (accelerated processing units).

To accelerate creating their catalog of server cartridges, HP is creating the HP Pathfinder Innovation Ecosystem, a business and technology development framework to accelerate their partners' time-to-market. HP will not only aim this program at chip manufacturers, it will also include OS and application developers. Server cartridges will be sold by HP as part of their product portfolio. It will enable smaller and niche accelerator vendors by exposing them to HP's customer base, will in return enable HP's customers by giving them wide choice in accelerator technologies.

Hardware Acceleration for Workloads and Applications

Most of the questions we receive on hyperscale infrastructure revolve around ARM. Loosely rank ordered by the frequency we have been asked:

- Which workloads is ARM good for in comparison to x86?
- Where will 32-bit ARM be used vs. 64-bit ARM?
- Is there a difference between ARM implementations?

If a service requires legacy x86-based applications code, there is no substitute. That is one of the primary reasons Intel's Atom is a good choice for many hosters.

Additionally the ARM community still lags Intel and AMD in raw performance per core, so if an application's threads are performance sensitive (and cannot easily be converted to a parallelized equivalent that can leverage more but slower cores), then those vendors are a good choice, though the application may not be sensitive to the x86 instruction set. ARM's licensees will eventually increase their performance and get close enough to Intel and AMD so that this will not be a substantial issue.

ARM core "bittedness" is a concern for some customers for a short window of time, and we are rapidly moving through that window. Most server runtime frameworks and applications are 64-bit today, but if a service provider has source code access or control then experimenting with and deploying 32-bit ARM applications in advance of a wide selection of 64-bit ARM processors launching into the market should not be a problem.

In general, any workload that parallelizes itself into threads that do not require high frequency cores is a good target for the ARM community. This is not limited to web serving and caching, there are even analytic workloads that fit this profile. Today most workloads make a trade-off from low 1GHz to low 2GHz core frequency ranges. I do not expect this to drift upwards significantly, as these workloads already scale nicely by adding more cores – the hyperscale data centers running them would rather have lower power and in many instances they would rather have better compute offload, too.

There are many ways ARM licensees differentiate their system on chip (SoC) designs. Some have architecture licenses and implement their own microcode engines – i.e. core designs. Many differentiate based on their on-chip system bus, some design their own and others license and implement various levels of ARM system bus designs. And then there are memory controllers, integrated Ethernet NICs and switches, etc. We believe there are no intrinsic system-level advantages evident by comparing chips and SoCs. The power efficient server SoC market could very much use a common set of relevant rack-level server benchmarks.

We are frequently asked about hyperscale and virtualization. A high-level answer is that merchant cloud and hosting services are most likely to implement virtualization. Many other services simply do not need to implement robust virtualization – it introduces unneeded layers of abstraction and complexity, plus it extracts nominal power and performance penalties. Virtualization is in many ways an antithesis to hyperscale computing. Some services elect to implement a lightweight form of virtualization to more facilely handle application crashes and server health monitoring.

The other interesting set of questions is about compute offload and acceleration:

- Does GPU compute have a place in servers aside from remote desktop rendering (includes game servers) and HPC?
- How do DSPs apply to server-side compute offload?
- What are the opportunities for specialized compute offload, in the form of FPGAs, custom offload engines, etc.?

Compute offload is beholden to Amdahl's Law. Offload engines have typically worked best when data can be streamed through them at source data rates with low latencies. In the past this has meant fixed-function acceleration worked best. One of the key issues with programmable compute offload – GPU, DSP, whatever – is that loading an application into dedicated offload memory takes time, loading a data set into offload memory takes time, and returning a resulting data set into a server node's main memory takes time. We believe that programmable compute offload engines of any type must be equal citizens to the processor cores with respect to their access to a server node's main memory, plus network and storage I/O resources. The HSA (Heterogeneous System Architecture) Foundation's community has a technically sound approach, and NVIDIA recently announced plans to do so as well.

When combined with ARM CPU cores, GPU or DSP compute offload may be able to help ARM-based SoCs offset x86 processors' native core speed advantage…at least for highly parallel tasks like image processing and malware detection, which is similar to network switching deep packet inspection.

At this early stage of hyperscale market evolution, an accelerator vendor's choice of using CPU, GPU, DSP, FPGA, dedicated logic, and/or combinations of them to accelerate a specific workload or application usually depends on which technologies they have already deployed in adjacent markets. AMD has GPUs, Intel and Tilera have interesting many-small-core architectures, Texas Instruments has DSPs, APM has integrated Ethernet controllers and a programmable microcontroller core, and Calxeda (focused entirely on the server market) has integrated Ethernet controllers and an integrated Ethernet switch.

APUs, which combine processor cores and compute offload engines into one SoC, should offer the best opportunity to balance CPU compute with specialized offload acceleration. APUs that enable their CPU cores and offload engines to share main memory will have an inherent advantage – they will save costs via simpler designs, non-redundant banks of memory, and the ability to pass pointers to memory locations instead of moving data will improve performance dramatically.

Not all of these vendors will compete for the same workloads, however, this variety will push the hyperscale infrastructure market to standardize benchmarks for high value workloads – there will be no obvious leaders until solid metrics are established.

HP's Moonshot 1500 System will provide a fascinating test bed for compute offload vendors to test their performance within a common system-level performance framework.

Conclusion

HP's new Moonshot 1500 System creates a standard set of infrastructure that hosts customizable but yet interchangeable server cartridges. Interfaces matter, and HP is creating a multivendor ecosystem around their Moonshot 1500 System interfaces, which they describe as a multiyear, multi-phased program – a long-term commitment.

HP's Moonshot 1500 System strategy is focused on specialization through modularity. HP's customers want to optimize their in-house applications, but do not know what that means yet for the mix of compute, storage, and networking on this new system. HP's customers can use the Moonshot 1500 System to right size their silicon and system infrastructure for their services and target applications.

In the early days, IT system manufacturers built their own proprietary processors. After Moore's Law and continuing integration enabled single chip processors, the IT industry decoupled the system from the processor instruction set and x86 became the de facto standard. HP is taking the next step by offering a modular system that lets customers choose the right architecture to optimize their service for specific workloads and their own in-house applications.

Author
Paul Teich, Analyst at Moor Insights & Strategy

Editor
Patrick Moorhead, President & Principal Analyst at Moor Insights & Strategy

Inquiries
Please contact us at the email address above if you would like to discuss this report and Moor Insights & Strategy will promptly respond.

Disclosures
Moor Insights & Strategy has a consulting relationship with HP. This paper was commissioned by HP. No employees at the firm hold any equity positions with HP.

HP Publication 4AA4-6179ENW

HP Moonshot:
www.hp.com/go/moonshot