Intel's Disaggregated Server Rack

Does "Disaggregation" Really Mean Anything?

There's been a lot of discussion about "disaggregated" servers, racks, and datacenters since Facebook and Intel promoted their vision for the phrase at the Open Compute Summit at the start of this year. Haven't we spent the last few decades disaggregating datacenter architecture? And if so, what does disaggregation mean now? Is it something different?

Executive Summary

The proposed Facebook + Intel disaggregated server rack is an extension of current system architectures introduced by AMD, Calxeda, and HP. While scale is an ambitious differentiator, it is a bit of useful misdirection for Intel with respect to their target markets and workloads.

The core differentiator for this interpretation of disaggregation is "atomicity" – separation of components at a functional level to enable a range of related workloads that require some scale as a baseline. Complete disaggregation is not a technology aimed at all datacenters and every workload, at least not in this decade and perhaps not ever.

Strictly speaking, "disaggregation" means dividing an integrated whole thing (an "aggregate") into its component parts.

Facebook and Intel are designing a specialized, completely disaggregated architecture optimized for large address space in-memory databases and analytics. This architecture is not intended to take over the entire datacenter. But it will be relevant for many classes of Big Data analytics, and it just might (finally) displace mainframes. IBM is reacting to this potential threat with OpenPOWER, a wildcard competitor for ARM's server licensees.

Hyperscale State-of-the-Art

Over the last decade, hyperscale datacenters have developed specialized architectures that partition workloads so that each runs on more optimal hardware. Specialized architectures and specialized workload accelerators reduce capital expenses by right-sizing compute, storage, and network resources to each workload; they also reduce power consumption and other operational expenses. As hyperscale customers push toward increasing workload optimization, they are spurring their component vendors to create both new network topologies and specialized hardware acceleration.

There are bottom-line economic benefits to storing data specific to a given workload logically close to the compute resources for that workload. Storage proximity counters the trend of consolidating storage into remote network attached storage (NAS) and storage area network (SAN) solutions. Storage server appliances address data locality

more efficiently by physically residing in the same cluster of racks as their workload compute resource counterparts. Data flow between compute servers and these storage server appliances is referred to as "east-west" flow. The goal for designing datacenter architecture at rack-scale is to minimize north-south flow by localizing compute and storage resources to a row or a small group of racks and enabling east-west bandwidth on that local scale.

Therefore, the networking goal for workload optimization is to create a localized east-west network fabric that reduces the need for north-south traffic as much as possible. This has created new product concepts, including Calxeda's EnergyCore architecture, AMD's SeaMicro Freedom Fabric, and HP's Moonshot System. All of these system architectures attempt to achieve a better balance of compute, storage, and networking resources at a rack-local level. They reduce hardware redundancy and north-south data flow via built-in east-west network fabrics.

East-west fabric design has, in turn, created a need for distributed object stores – modern reinventions of a file system, designed from the start to span networks and to be implemented with inexpensive media that will frequently fail. Distributed object stores use north-south network bandwidth sparingly. They automatically try to move data as close to compute resources as possible, and in a timely manner. And they do all of the other things we expect of file systems: enable backups, ensure availability of data when needed, enforce data integrity, etc.

Compass Headings

Hierarchical datacenter networks are designed to transport data from a leaf node, a server, up to a datacenter backbone network consisting of "core" switches, and then back down to another leaf node, such as another server or a gateway to a different network. Between leaf nodes and core switches are layers of intermediate switches that aggregate increasing amounts of traffic as data flows from leaf nodes to the core switches (defined as "northbound" flow), and those intermediate switches also disaggregate traffic flowing from the core routers to the leaf nodes ("southbound"). Intricate multi-level hierarchical network architectures have been created to handle varying amounts of north-south data flow.

Virtualized Storage

Virtual storage emerged to address challenges in data synchronization, integrity and availability. Federations of networked servers were deployed to address increasingly transactional web workloads, creating a challenge for timely and reliable sharing of data between server nodes. As a result, the amount of direct attached storage (DAS) in each server node has declined as shared storage was pushed out to SAN and NAS solutions. But SAN and NAS had an unintended consequence – they increased north-south data flow traffic as virtual storage became, in effect, another kind of server. This happened at the same time Internet and web traffic increased dramatically, compounding data flow demands on hierarchical datacenter network architectures.

The emergence of Big Data usage models has introduced even more variation into storage system design: the concept of "cold" or "dark" storage that is powered-down when not in use, but can be rapidly powered back up when data is needed. Facebook has perhaps the most public example of cold storage in their photo archives. Old Facebook photographs (over several weeks old) are almost never accessed again, but they must be rapidly available for access, on a scale of many seconds (but Facebook can't keep its customers waiting too long…15 seconds borders on "too long"). Tape is not fast enough, but spinning media burns power and solid-state drives (SSD) are too expensive. Facebook's answer is to spin-down hard disk drives (HDD) within a minimum-cost but reliable "cold storage" rack-level design, and then spin them up when needed. The photos are accessible if or when customers look for them.

What Does "Disaggregate" Mean?

Strictly speaking, to "disaggregate" means to divide an integrated whole thing (an "aggregate") into its component parts. What is it that we're "disaggregating" as we move from datacenters designed to serve enterprise IT workloads to hyperscale datacenters designed to serve Internet-scale workloads?

A modern IT datacenter is already disaggregated when compared to self-contained solutions such as mainframes. Compute is separate from storage is separate from network. They are all different products, with different dominant designs, and manufactured mostly by different sets of vendors (compute is dominated by HP and Dell, networking by Cisco and Juniper, and storage by EMC and NetApp).

In many respects, Calxeda, AMD, and HP are "re-aggregating" computer architecture. They are packing compute, storage, and local network fabric into a more tightly-integrated rack-level architecture, optimizing east-west data flow at a local level. Instead of carving datacenter architecture into its component pieces, they are throwing components into a blender for a more fine-grained approach to optimizing hardware for specific workloads.

And, as traditional storage appliances absorb more compute capability to become distributed object storage server appliances and network switches and routers absorb more compute capability to become software defined networking (SDN) enabled, we expect "re-aggregation" possibilities and experiments to expand.

As of the beginning of this year, most of the east-west discussion and product innovation had taken place inside of a rack-mounted chassis (AMD at 10U chassis height, HP at 4.3U, and Calxeda at 4U) with today's hierarchical networks connecting those rack-mounted chassis.

Intel, Facebook, OCP, and Disaggregation

On January 16, 2013, Intel and Facebook announced a collaboration to define next-generation "disaggregated, rack-scale server" architecture and designs (for simplicity, we'll refer to this as "OCP DRS"). An executive-level overview is posted on the Open Compute Project's Summit site.

The OCP DRS disaggregation goal is to separate compute and storage within a rack. But as mentioned above, those resources are already separate. What makes this proposal different?

The primary differentiator is atomicity. OCP DRS isn't referring to separation of logical compute, storage, and networking resources. OCP DRS atomicity means the separation of hardware component classes at a hardware-defined functional level and interconnection of those components with separate distribute switching functions. It is an audacious thought.

Instead of distributed storage servers and compute nodes, a future OCP DRS rack design might have a few trays of processors, a few trays of system memory, lots of trays containing a mix of SSDs and HDDs, a distributed east-west fabric to tie all of those trays together, and a single consolidated datacenter network interface for north-south data traffic. That consolidated network interface would take the place of today's top of rack (TOR) or end of rack (EOR) switches.

The central, enabling feature of the OCP DRS proposal is its high-speed east-west network fabric and switching capability. Current datacenter architecture is built on layers of data transport technologies, starting with chip-level bus architectures. There is a tension between physical distance and system performance, driven by the cost to deliver low latency and high bandwidth access to data across a given physical distance.

OCP DRS highlights this core tension as does no other fabric architecture outside of niche high performance computing (HPC) applications.

Table 1: OCP DRS Compared to Current Designs

	Current Designs	OCP DRS
Design Point	Collection of Modular System Chassis	Single Rack is Smallest Unit, Spans Many Racks
Atomicity	System Chassis within Rack	Functional Components
Interconnect	Minimize Costs within Standard Topologies	Create New Topologies to Enable Higher-Order Efficiencies

The big nut to crack for OCP DRS is to deliver required switching performance at reasonable cost in a competitive timeframe, which will then enable the promised efficiencies of scale within the capital and operating expense benefits (CAPEX and OPEX, respectively) projected for large scale datacenter deployments.

Every bit of this proposal is predicated on Intel's ability to deliver on its promised silicon photonics technologies – within a competitive schedule and with competitive performance, cost and scalability. Practically speaking, the OCP DRS cannot be deployed in volume until the second half of this decade, which, of course, starts in less than 18 months.

Copper Interconnects vs. Intel's Silicon Photonics

There is a tension between distance and interconnect performance for copper-based (electrical, not optical) networking technologies.

Table 2: Copper interconnect standards for the middle of this decade.

Interconnect	Connects	Bandwidth	Max Distance
JEDEC	Off-Chip Memory	**DDR4** x16 Lanes Max ~8.5 GB/s x1 channel Max ~34 GB/s x4 channels	Typically <250mm
PCI-Express (PCIe)	Off-Chip I/O	Per Lane, x1 to x16 Lanes: **3.0** = 985 MB/s **4.0** = 1.97 GB/s Max ~32 GB/s for 4.0 x16	Typically <250mm
SATA	Board to HDD/SSD	**3.0** = 600 MB/s	1m
Fibre Channel	System to Disk Array	400 MB/s	12m
Ethernet 10 Gbps	System to Switch, Switch to Core	**802.3an** = ~1.2 GB/s	100m
Ethernet 40-100 Gbps	Core to Core	**802.3ba** = 5 to 12 GB/s	100m

In general, latencies become longer at each step from the top of the table to the bottom, combining with slower bandwidth to produce much slower aggregate performance. We also assume that all of these are used as point-to-point lanes or cables within a chassis or within a datacenter, and that at least one end of the connection terminates at some sort of switch.

It is very hard to drive digital signals across copper wires faster than we can already achieve. The above technologies pay a price to achieve their data rates – they either use multiple lanes to get faster data rates or they use more expensive and higher power signal drivers and cabling (also coupled with multiple lanes) to achieve the same effect.

Intel has proposed that OCP DRS be built on top of Intel's silicon photonics technologies. Their website says that they reached ~6 GB/s by July 2010 and their slides from January's OCP Summit say that they have samples running at twice that, on the order of 12 GB/s, on a single mode of a single fiber. At Intel's Data Center Day 22 July 2103, Intel disclosed more detail about their current rack-level capabilities, including a silicon photonics "patch panel" implementing 25 Gbps southbound and 100 Gbps northbound interconnect bandwidth.

Figure 1: Maximum Interconnect Bandwidth vs. Maximum Interconnect Length (Log-Log Chart)

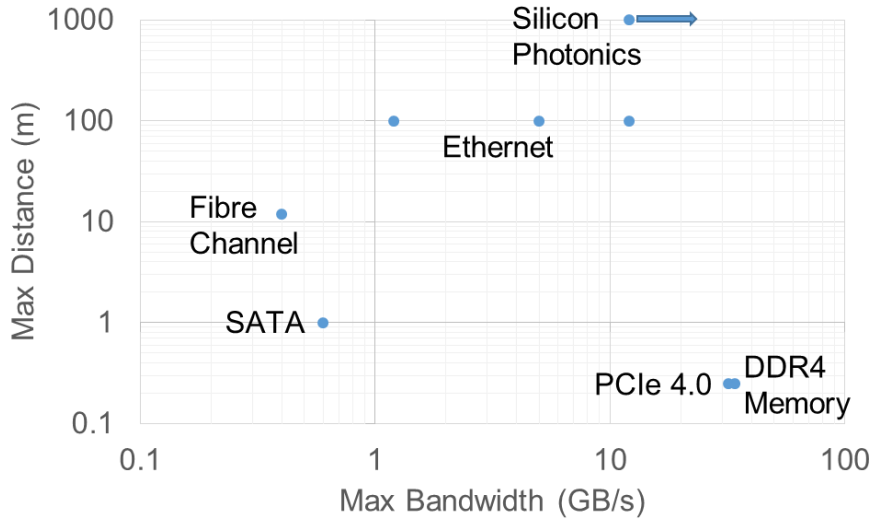
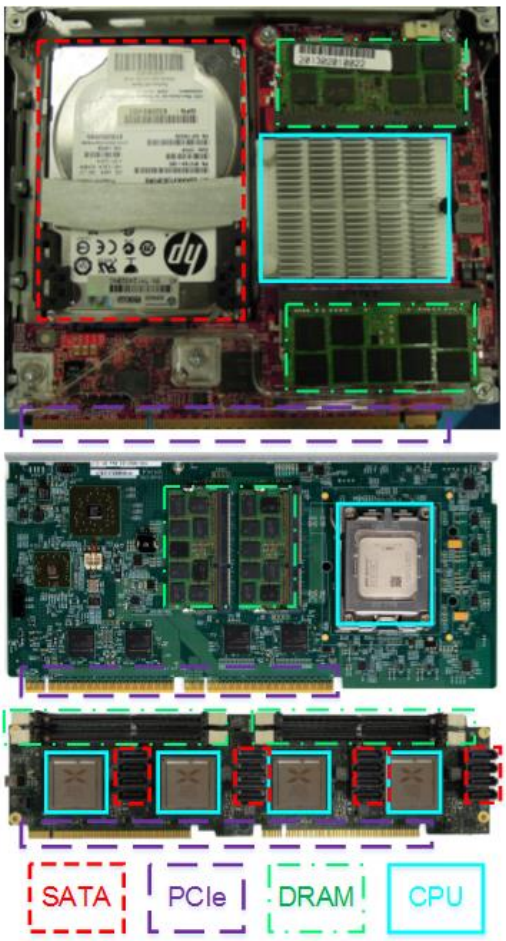


Figure 2: Physical Board Layouts



HP Moonshot Cartridge with Intel Atom processor

- Single processor on board
- Two memory boards
- SATA connected HDD on board
- Ethernet to Moonshot Ethernet switch and PCIe lanes for other functions

AMD SeaMicro processor card with AMD Opteron processor

- Single processor on board
- Two memory boards
- SATA and Ethernet virtualized across 3D torus interconnect via PCIe lanes

Calxeda EnergyCard with EnergyCore SoCs

- Four SoCs on board
- One dedicated memory card
- Four dedicated SATA ports
- Ethernet and PCIe signaling to system interconnect

It is reasonable to expect that in 18-24 months Intel will have productized technologies (silicon transceivers, cable connectors and switches) in the 12 to 24 GB/s bandwidth range for single mode, single fiber cables. It is also reasonable to expect that they will be able to obtain transmission distances of much greater than 100m at maximum bandwidth, as optical fiber has very good transmission distance.

This progression of silicon photonics bandwidth increases implies that, in the 2016 timeframe, Intel's silicon photonics technologies should be bandwidth-competitive with 40 to 100 Gbps Ethernet. Perhaps Intel's silicon photonics cables and transceivers might also be less expensive than their contemporary Ethernet equivalents. It also implies that silicon photonics should be competitive with short-distance copper interconnect technologies using a single fiber and mode by the end of the decade.

System Architecture Implications

In today's commodity server designs, all memory is local to a processor chip. There are no contiguous and large pools of system memory that span compute chasses. Each processor in a multiprocessor design has local memory, with rare exceptions when a 2P or 4P design does not implement physical memory connectors for one socket in a 2P design (the vast majority of commodity servers shipped) or some sockets in a 4P or higher design. This still holds true for the above mentioned fabric-based products.

Using current modular chassis system design with the above data rates, each server processor chip in the 2016 timeframe might have up to 34 GB/s of dedicated, very low latency bandwidth to its memory. Its memory is (in general and as mentioned above) not shared with more than one other processor chip. Most of the previously-mentioned hyperscale experiments in east-west localization point to a future in which, for many workloads, processor chips do not share memory with each other at all.

PCIe bandwidth to local network and I/O resources (including storage) becomes a little more interesting. Each processor can have tens of GB/s of dedicated bandwidth, but new east-west fabrics potentially limit the number of effective traffic lanes in a mesh or other local fabric topologies.

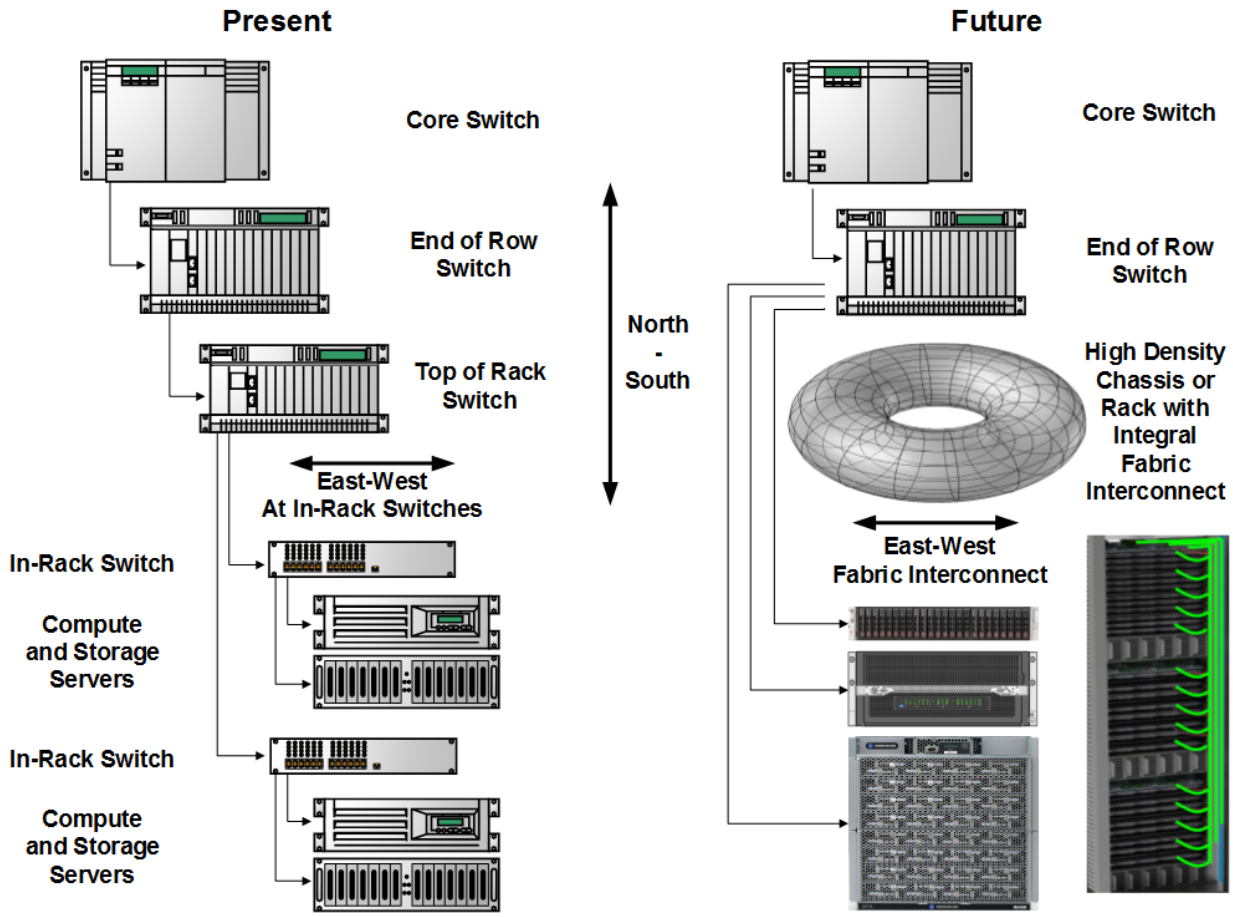
In a single-chip server SoC product, PCIe may not be exposed outside of a chip at all, in favor of direct interconnects to the physical east-west fabric. Many chip manufacturers are considering stacking memory die in future SoC products, but until that happens they will all incorporate an external interface to their dedicated system memory.

Every SoC manufacturer is investigating die-stacking or other memory integration technologies, because that integration will drive mobile and Internet of Things power usage, product size, and economies of scale in this same time period. These SoC in-package memories will act as large local caches within an OCP DRS architecture.

The predominant datacenter architecture is that many servers in a rack are aggregated through an in-rack switch (to aid east-west traffic flow), and then those switches

aggregate through a TOR switch. The TOR switches in a row of racks all aggregate through an EOR switch, and the EOR switches link to the datacenter core switches.

Figure 3: Network and Fabric Topologies



East-west fabric vendors hope to simplify this architecture by creating self-contained network fabrics that will make in-rack switches redundant and perhaps do the same for TOR switches.

The challenge for OCP DRS is that it will require new architectural design and bundles of optical cables to provide enough bandwidth to physically separate components that are today quite close together.

The easiest way to start would be to keep processors and their memory collocated and only separate compute nodes from storage and network resources.

This is what AMD, Calxeda, and HP have ***already*** implemented, albeit at chassis scale and not at rack scale. Generically, each of these architectures packs processor plus memory "server nodes" onto high density compute cards, with network and storage resources virtualized across in-chassis local east-west network fabrics.

The OCP DRS gets rid of the chassis enclosure and supersizes the east-west network fabric to an entire rack and then to a group of racks. This isn't really a difference in architecture as much as a simple difference in scale. And if this were all that OCP DRS wanted to accomplish for many high volume workloads, they don't have to wait for silicon photonics to implement this simple scaling exercise. Intel could implement a rack-scale architecture that builds on current technologies now, with 10 Gbps copper Ethernet (Intel says that OCP DRS is not their only architectural direction for hyperscale). And, Facebook's own Open Vault "Knox" storage server architecture follows a completely different architectural direction.

Which Workloads Require Complete Disaggregation?

Physically separating main memory from processor cores requires new memory architectures, and more than anything else it requires a huge leap in available network bandwidth to bridge that physical distance.

As you can see in Table 2, PCIe 4.0 will provide roughly equivalent bandwidth to multiple channels of fast DDR4 memory in the middle of this decade. This is the first time that a standard, widely available I/O interface will be able to provide competitive bandwidth against multiple channels of contemporary state-of-the-art system memory.

The tricky part will be to ensure that PCIe latencies are not a performance show-stopper. Latencies can be reduced by:

- Minimizing the number of network switch "hops" between processor and memory
- Ensuring "non-blocking" bandwidth and fast switching times at each switch
- Maximizing signal integrity and minimizing communications errors over each point-to-point network hop

When PCIe 4.0 bandwidth is available, then Intel might:

- Build a processor tray with enough PCIe switching capability to access main memory remotely
- Build matching processor tray and memory tray switches, with enough aggregate memory bandwidth between them to serve a given workload
- Build a memory tray (really a rack level architecture spanning multiple memory trays) with memory address space aware switching capability and enough memory bandwidth in the switch to enable all of the memory in the tray to perform at stated speeds

This will be expensive. It will require bundles of fiber optic cabling to implement (three 100 Gbps optical fiber interconnects to match memory bandwidth to one processor SoC). Silicon photonics will help reduce these attributes of their new east-west fabric:

- Power consumption
- Board area and rack volume
- Cost of network transceivers

However, Intel will have to manufacture an entirely new class of switch architectures to support the switching speeds, non-blocking bandwidth, and aggregate throughput required to serve target workloads. Those switches will probably not be price competitive with commodity Ethernet switches.

We believe that Intel is building a silicon photonics enabled system architecture optimized for OCP DRS large address space in-memory databases and analytics. One of the high value markets such an architecture would open for them would be large transactional systems – i.e. mainframe replacements.

The bill of materials profile looks like it will appropriate for the high value applications they will target. Facebook will be the first software-as-a-service (SaaS) vendor on their block to take advantage of new analytics economies of scale, including Intel's investment in open source analytic frameworks like Hadoop. It should be a very happy product development arrangement for both Facebook and Intel.

For mainframe replacements, Intel has spent a decade honing their software migration tools on Itanium workload migration projects. While x86-64 blunted much of Itanium's market potential, Intel has built a strong experience base in automatic translation and migration of mission critical enterprise IT workloads.

Implications

Moor Insights & Strategy is an advocate of specialized workload acceleration, from single SoC smart objects and mobile computing up through hyperscale datacenter architectures. There will not be one "correct" architectural answer for every device and workload. OCP DRS will address a significant and profitable at-scale Big Data market segment, but it is not a viable answer for a broad range of workloads through the rest of this decade.

We also believe that IBM's recent disclosure of their OpenPOWER initiative is a direct result of Intel's OCP DRS design potential as a mainframe replacement, and not necessarily a response to ARM server licensees. ARM licensees do not pose a direct threat to IBM's core hardware and software business. IBM needs partners to invest in its POWER architecture to keep pace with Intel. To do so, IBM had to create their new OpenPOWER ecosystem to address this high-end threat as IBM's former embedded partners move away from POWER to architectures like ARM.

Intel, on the other hand, presents an entirely different R&D and market potential for delivering a new architecture for mission critical IT workloads. Intel has experience in core enterprise markets and has demonstrated long-term investment tenacity in the face of overwhelming odds with their Itanium product line.

We believe that IBM was forced into rapidly forming and announcing the OpenPOWER consortium once they comprehended OCP DRS and both its Big Data analytics and its mainframe replacement potential. IBM's new OpenPOWER consortium poses a

problem for ARM's server licensees. IBM's opportunity is to become the credible enterprise-worthy, vertically-integrated Big Data analytics alternative to Intel, and to relegate ARM to server appliances and newer services-oriented workloads.

The ARM server community has been fairly quiet as they prepare their 64-bit offerings for late this year and throughout 2014. We expect IBM to take advantage of ARM's press and marketing silence and try to steal some of ARM's server mind-share over the next few quarters.

There is a cross-cultural ancient curse that translates to this English phrase: "may you live in interesting times." The datacenter industry is in a very interesting period of rapid transition to new workloads and to new datacenter technologies optimized and tuned to serve those new workloads. "Interesting" is an understatement!

Important Information About This Paper

Author
Paul Teich, Senior Analyst at Moor Insights & Strategy.

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

Inquiries
Please contact us here if you would like to discuss this report and Moor Insights & Strategy will promptly respond.

Citations
This note or paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title and "Moor Insights & Strategy". Non-press and non-analysts must receive prior written permission by Moor Insights & Strategy for any citations.

Licensing

Disclosures
Moor Insights & Strategy provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this documented.

DISCLAIMER
The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness or adequacy of such information and shall have no liability for errors, omissions or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.

