How to Manage One Trillion Devices on the Edge

New Compute Models Require New Thinking

The Edge, the Datacenter, and New Design Principles

The world of compute is changing rapidly, as is the traditional view of a physical building, or buildings filled with servers, storage, and networking to "run the business". Cloud computing, distributed cloud computing, and edge computing will all be fed by a 5G access network, forcing IT organizations to think and plan differently.

The environments managed by the IT organizations of tomorrow will extend beyond the traditional datacenter into the distributed cloud and edge environments. Ruggedized servers sitting on oil rigs; cameras, sensors, and micro datacenters powering smart cities; even mobilized healthcare platforms that provide patient care and monitoring are all examples of why datacenter assets must exist on the edge for petabytes and zettabytes of data to be transformed, analyzed, and turned into actionable intelligence in real-time.

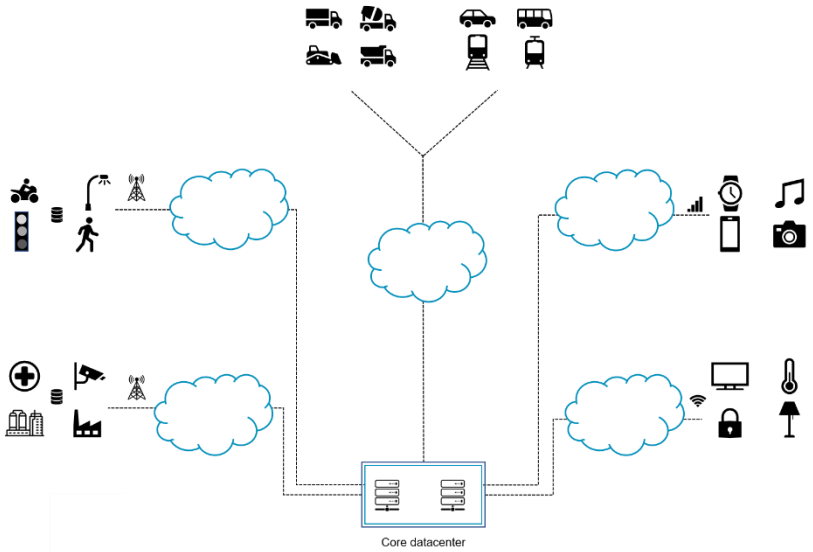
Arm predicts the world will see up to one trillion edge-connected devices by 2035.[1] While this number is staggering, a quick look at the underlying dynamics supports this predicted magnitude of devices. Consider the following:

- Consumer adoption of Internet of Things (IoT) is forecasted to continue to skyrocket. In North America, the average consumer will own 13 connected devices by 2021.
- Connected cars and autonomous vehicles that move people and goods will number in the hundreds of millions and each vehicle will require thousands of sensors.
- Artificial Intelligence (AI) will drive everything from farming to manufacturing to healthcare, requiring hundreds of billions of sensors globally.
- Smart cities around the globe will be populated with billions of high definition cameras connected with hundreds of billions of sensors and devices that

[1] https://pages.arm.com/route-to-trillion-devices.html

manage traffic flow, respond to emergencies, and even power on/power off street lamps for pedestrian traffic.

FIGURE 1: THE WORLD OF ONE TRILLION DEVICES



Source: Moor Insights & Strategy

This explosion of devices and data presents several challenges for organizations which include:

Compute must be local to the data as needed to provide real-time ingest, transformation, and analysis. Otherwise, the latency associated with edge-to-core round trip would render actionable intelligence outdated and the sheer volume of data generated would overwhelm available bandwidth. Compute must also be as performant as possible to adequately process the expected quantities of data in real-time at the edge or even greater quantities of aggregated data in the core datacenter.

Networking must be smart, flexible, and robust enough to move data across the edge, the distributed cloud, and the core datacenter.

The core datacenter must be designed in a way that fully enables the core infrastructure needs of the enterprise and seamlessly support the data influx and compute requirements on the edge.

The software community has responded to the needs of both the distributed cloud and the edge by embracing open source, service-oriented architectures (SOA), and

microservices. The resulting highly-distributed applications that are emerging have outpaced the natural evolution of server hardware design. To stay relevant, the workloads and software applications driving organizations of all sizes must run on the range of servers, appliances, and devices that make up the modern, distributed datacenter.

For example, managing smart cities requires localized compute (coupled with high speed networking) to analyze and re-route traffic in the event of an accident in real-time, and it must continue to process critical data even if the network connecting to the core were to go down. This example alone requires a range of hardware in terms of performance, power, and form factor since the physical, power, and thermal requirements for each use case are unique. Conversely, the hardware requirements to support an oil refinery plant have an entirely different set of requirements. While these conflicting requirements have long existed, the availability of software stacks that drive greater control and efficiency require a software and hardware ecosystem where choice enables necessary tailoring.

The confluence of these dynamics requires a shift in datacenter design principles. Whereas datacenter architects have traditionally taken a “datacenter out” approach to design, they will now have to consider designing from the edge to the core datacenter as most data will be generated, transformed, and acted upon at the edge.

ARM – SCALABILITY AND SECURITY FROM THE EDGE TO THE CORE DATACENTER

Despite the new “edge to datacenter” design approach, the fundamental constraints of datacenter design will not change: deliver the best performance within a power budget in a secure operating environment with high bandwidth networking to support the business and its users at the lowest possible total cost of ownership (TCO). Given the explosion of the Industrial Internet of Things (IIoT) and edge computing, Moor Insights & Strategy (MI&S) views these as perhaps more important than ever. Three key tenets of datacenter design – performance, power, and security – all must start in silicon with:

- Performance of infrastructure that is both scalable and predictable as resource demands grow and shrink with the demands of workloads that power the edge, the cloud, and the enterprise datacenter.
- Power profiles that support the extended datacenter. Low power profiles (as low as sub 10W) may be required for the equipment that physically sits on (and drives) a manufacturing floor or must operate in a fan-less environment. Mid-

range power profiles are needed for the mobile edge compute environments where inference and processing is required, and higher power profiles that drive higher end servers in a traditional datacenter environment.
- Security that can span the physical device to the data generated, analyzed, moved, and stored.

The usefulness of silicon is only as strong as the richness of its ecosystem which is made up of the independent hardware vendors (IHV) and independent software vendors (ISVs) that build targeted solutions on open-source software (OSS) running on silicon intellectual property (IP). While today's datacenter can be considered an x86 world, it is a combination of general-purpose compute servers in combination with purpose-built platforms for networking, storage, and other functions. As the datacenter extends to the edge and beyond, diversity of architectures that include silicon, system, and software will be the norm.

When considering this "edge-to-datacenter" approach to design, finding a software architecture that can scale and move from the edge to the datacenter is critical since infrastructure built on different silicon products and product types will be required. For example, the convergence of traditional IT operational technology (OT – the organization and technology that drives industrial networks) will require a software architecture that can connect devices and the data they generate with industry standard servers. In addition, the OT has to be able to perform real-time analysis in managing that industrial environment. Otherwise, the promise of operational efficiency is lost.

Arm is well-positioned to support the needs of the modern datacenter for a few reasons:

- Arm IP has a large footprint across the billions of devices that make up the edge, the network, the distributed cloud, and the core datacenter today. This footprint will increase as we near the trillion devices of tomorrow.
- Arm IP powers devices (such as smart Network Interface Cards or smart NICs) and servers that scale, spanning fan-less, very low power (<15W) and high-end hyperscale (>100 cores operating in >150W TDP range).
- The power, performance, and security profile in Arm IP is consistent with the device to the core datacenter approach.
- The growing presence of open source software and its disaggregation from hardware, along with the growing integration of edge to datacenter, leads solutions providers to look for the right fit instead of a common instruction set architecture.

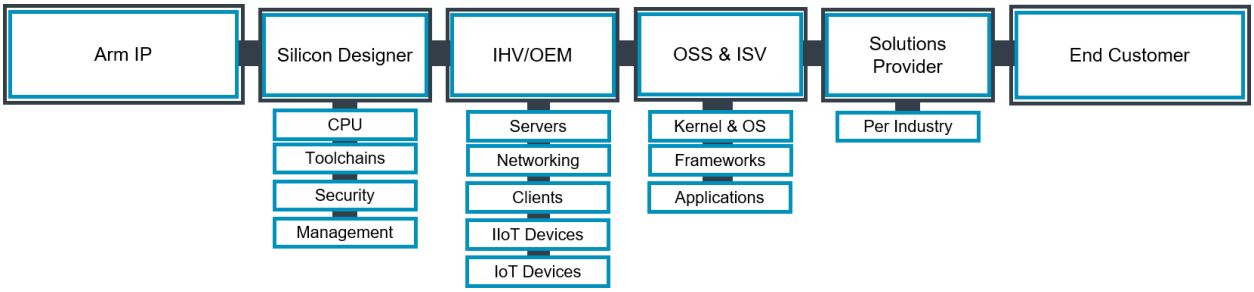
- Perhaps most importantly, the breadth of the Arm ecosystem is filled with partners collaborating to design and develop both general purpose and tailored solutions that leverage Arm IP across vertical industries, devices, and usage models.

IT'S ALL ABOUT THE ECOSYSTEM

As previously mentioned, a common architecture inclusive of all the elements that make up the modern datacenter seems difficult to achieve, if not impossible. Yet, MI&S believes this is what makes Arm unique.

Arm is also unique in that it is a provider of silicon IP, not a manufacturer of processors. One of the benefits of being an IP provider is that it becomes the engine of innovation for technology companies developing industry-specific system-on-chips (SoCs) and solutions. Arm being an IP provider allows tailoring per workload, per use case, per form factor, etc.

FIGURE 2 – THE ARM ECOSYSTEM



Source: Moor Insights & Strategy

Arm's model of licensing IP attracts an ecosystem of technology providers that can develop market specific SoCs and solutions for new workloads at a rapid pace and, as previously noted, with the right amount of resource for optimal performance. Arm's model extends this capability to OEMs seeking customized solutions for their specific use cases and its ecosystem offers the choice of silicon vendor while maintaining commonality of software architecture. IHVs looking to develop end-to-end solutions for a specific vertical can design around different central processing unit (CPU) silicon (or even design their own) or build custom application specific integrated circuits (ASICs) that are best fit for use with assurance of software compatibility from the device to the server.

For those that think of Arm as being the IP that powers embedded systems such as traditional network routers and switches, appliances, and the like – this is correct. Their success in networking and the cloud gives Arm a significant datacenter footprint. In fact, Arm is a significant player in the adoption of virtualized and containerized network function virtualization (NFV) in the telecommunications space. Operators can improve their agility and responsiveness to market needs and their ability to innovate quickly by moving from traditional network appliances with proprietary software to an open standards-based approach.

The Arm ecosystem increasingly finds traction in traditional datacenters and cloud providers. The Arm silicon ecosystem for server class CPUs and SoCs is rich with partners demonstrating performance that can support both traditional infrastructure needs, high performance computing (HPC) workloads, and AI.

Finally, the strength of the Arm ecosystem can be seen in the enterprise software that runs organizations. Support for Arm in the Linux Kernel led to support for all the major server operating system (OS) distributions including Red Hat, CentOS, Canonical, Ubuntu, etc. Moreover, VMware's recent announcement and demonstration of support for Arm in its ESXi hypervisor is perhaps the most profound proof point to the relevance of Arm in the datacenter.

DATA IS THE NEW CURRENCY AND CONNECTIVITY IS THE BANKING

Data drives the modern business, network, and datacenter. The separation between winners and losers in the digital economy will be measured in speed. How quickly can an organization collect, shape, analyze, and act on terabytes worth of data in real-time? How secure is an organization's data and IP - from raw ingest to analysis to cold store?

Ironically, in a market where software-defined everything is driving mindshare in the market, hardware has never been more important. While hardware's reliance on silicon may seem obvious, this statement has never been more true.

Consider the usage model where real-time data is collected from multiple assembly lines on a manufacturing floor and compared against multiple data sources to ensure the highest levels of utilization across all assembly lines. In this use case, the ability to design from the device to edge to network to core datacenter on a single underlying architecture can help organizations gain the performance, scale, connectivity, and security necessary to deliver that outcome.

IN CLOSING

The datacenter is no longer a physical construct but serves as centers of data and compute that reside in the cloud, on the edge, and in the wild. The datacenter is the raw data and intelligence that drive the modern business. It is the infrastructure that collects, moves, secures, and turns that data into action. In the modern datacenter, the lines between compute, networking, edge, cloud, device, and server are blurred and the path to one trillion connected devices further blurs any lines of demarcation.

Due to this, organizations must consider distributed, connected datacenter design methods that are flipped, moving from “edge to datacenter” instead of from the traditional “on premise” datacenter outward. As with any disruptive market, players will emerge that carry significant relevance. MI&S believes Arm is one of those rapidly emerging forces in this market. The company’s IP portfolio, breadth and depth of ecosystem, and footprint in the IoT, IIoT, and networking markets combined with its growing presence in the traditional and emerging datacenter spaces position it for success.

With Arm, there is a benefit for all players in the ecosystem. Silicon vendors can develop CPUs and chips that are both general purpose and tailored for specific workloads, scaling from devices to networking gear to datacenter infrastructure. Tier one operators can gain greater savings through vertically integrating their silicon supply chain. Hardware vendors and OEMs can quickly react to market needs because of the rapid pace of innovation. At the same time, software vendors can more cohesively design and develop end-to-end solutions which can be vertically tailored to industry-specific solutions providers. In the end, the customer should benefit the most.

IMPORTANT INFORMATION ABOUT THIS PAPER

CONTRIBUTOR
Matt Kimball, Senior Analyst at Moor Insights & Strategy

PUBLISHER
Patrick Moorhead, President, Principal Analyst

INQUIRIES
Contact us if you would like to discuss this report, and Moor Insights & Strategy will respond promptly.

CITATIONS
This 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
This paper was commissioned by Arm. 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 document.

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.