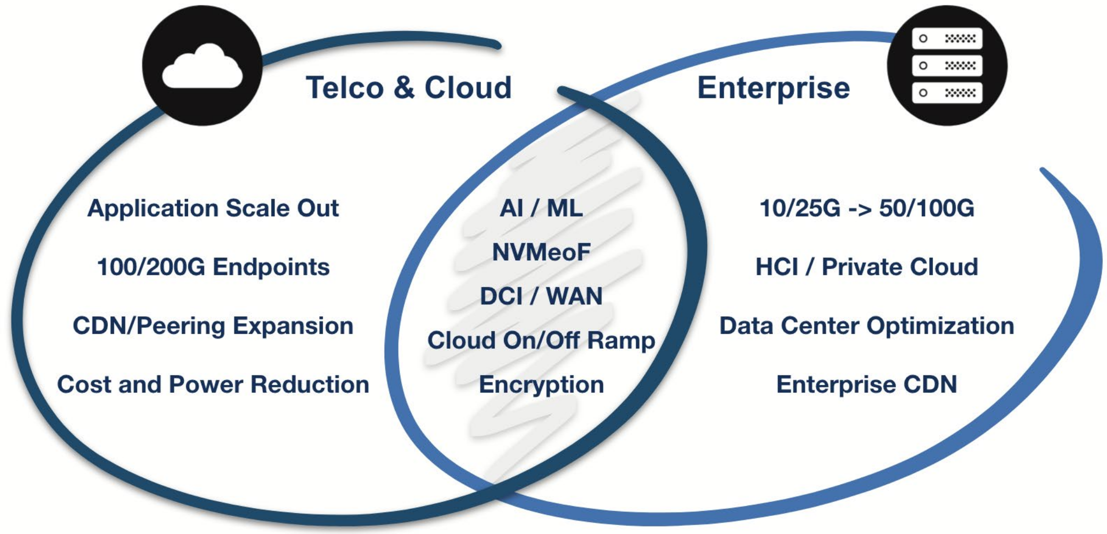
Simplifying 400G for Data Centers



Introduction

From its origins as an ultra-high performance technology, reserved for a few organizations with extreme networking demands, 400 Gigabit Ethernet (400GbE) has evolved into a mainstream technology providing significant advantages to network operators across a wide spectrum of use cases.

Both emerging requirements and strategic directions have converged driven by significant leaps in technology innovation to create an ideal environment for the adoption of 400GbE; Surprisingly, the motivation isn't always simply more bandwidth, the mainstream availability of 400GbE also presents opportunities for significant cost savings and efficiency improvements for organizations planning upgrade cycles or greenfield deployments.

In this paper we review some of the key benefits of 400GbE, made universally accessible through Arista's broad portfolio of platforms, each designed to fit different workload types and scale requirements.

Bandwidth

The most obvious attribute of 400GbE is bandwidth, with demand accelerating across a number of use cases in Cloud, Service Provider and Enterprise environments.

Internet, Media and 5G

At one end of the scale, high bandwidth applications such as rich media present a clear case for an increase in end to end capacity. Streaming video content now represents the overwhelming majority (80%+) of Internet traffic, with significant numbers of consumers switching from traditional broadcast to Internet streamed content. This unprecedented demand drives the expansion of Content Delivery Networks (CDN) with long-haul 400GbE and an upgrade of Internet Exchange Points (IXP) to support 400G peering between CDNs and ISP networks to reduce the number of peering cross-connects and allow for consolidation.

As mobile networks transition to 5G offering higher bandwidth and new applications they also deploy subscriber functions in disaggregated and distributed data centers (the new 5G edge). Increasing demand for video on a growing number of portable devices requires an edge content distribution model. In parallel, the growing number of devices, connected vehicles and IoT products drives mobile networks to support similar capabilities to their terrestrial counterparts.

The accelerating adoption of IP in broadcast media and production industries has generated higher demand for bandwidth. With raw media quality improving from 4K to 8K and beyond, device interface speeds have increased to 50 and 100G. New distributed production workflows requiring uncompressed video mean that large volumes of data must be transported between the source: film and TV studios, outside broadcast and film locations to the centralized destination of data centers and the cloud placing heavy loads on private and public infrastructures.

High density 400GbE routing solutions with scalable resources and advanced buffer management, such as Arista's 7280R3, 7500R3 and 7800R3 families have made Internet scale routing with comprehensive Service Provider functionality accessible.

Support for the innovative OSFP transceiver specification ensures the broadest support for higher power, forwards compatible 400G transceivers including Arista's innovative OSFP Line System (OSFP-LS). The OSFP-LS makes it possible to combine multiple 400G-ZR circuits onto a single fiber pair, interconnecting data centers and points of presence (POPs) at multi-terabit speeds at a fraction of the cost of traditional optical line systems.

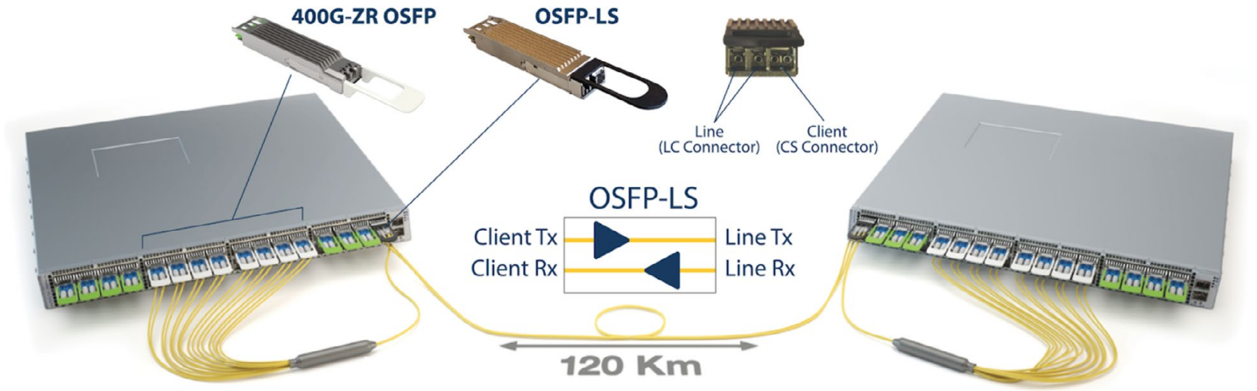


Figure 1. 400G-ZR with Arista's OSFP-LS providing 3.2Tbps over 120km dark fiber

High Performance Computing & Cloud Native Applications

High performance computing (HPC) with artificial intelligence/machine learning (AI/ML) are increasingly mainstream applications at the forefront of innovation in the use of automation, modelling and autonomous systems for research, financial services, manufacturing industries and in the broad commercial world. These next generation applications leverage ever larger data sets and increasing numbers of clustered compute nodes communicating in east-west patterns at high speed and low latency to operate effectively.

In addition to the need to build increasingly large clusters to support these applications, the adoption of FPGA and DPU based network adapters (SmartNICs) based on PCIe Gen4 for higher bandwidth, lower latency and higher throughput flash based storage systems and NVMe for distributed file systems, leveraging Remote Direct Memory Access (RDMA) has rapidly driven the network connectivity on servers from 10/25Gbps to 50G, 100G and 200 Gbps for the latest generation. Arista's 7050X4 and 7060X4 families are ideally suited to these applications offering a range of interfaces from 25G to 200G and 400G.

The 400G OSFP and QSFP-DD standards used to deliver 400GbE are both natively designed to allow a 400GbE port to support multiple other speeds with breakout, parallel media and previous generation form factors. The range of speeds include a single port with up to 8 x 50GbE, 4 x 100GbE or 2 x 200GbE connections. As a result a single, 1RU 400GbE switch with 32 OSFP ports could support a cluster of up to 256 x 50GbE nodes or act as a leaf device in a two-tier cluster supporting thousands of nodes.

The diagram below shows an example of a non-blocking 102.4Tbps cluster with 128 single-attached 50G compute nodes connecting to each leaf switch. Each leaf switch connects at 2 x 400G to each of eight spine switches yielding a total cluster size of 2048 compute nodes with no network oversubscription and only 24RU of network equipment.

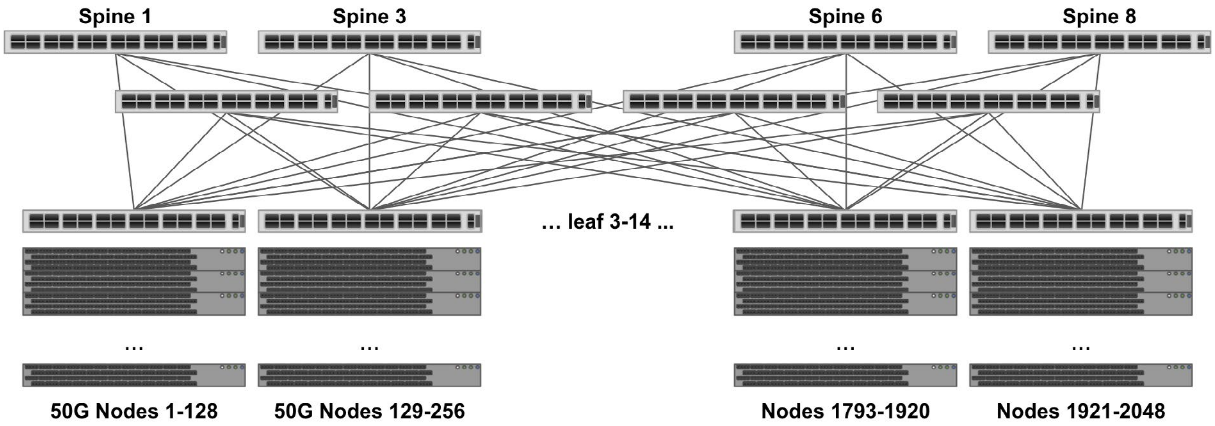


Figure 2. A non-blocking 102.4Tbps cluster of 50G attached servers using an 8-way leaf-spine topology

Enterprise Workloads

Enterprise workloads and applications themselves have also become more distributed, often located in dedicated facilities, independent from the workplace, and split between multiple on-premise data centers and public cloud facilities. This creates two emerging needs, addressed by 400GbE technologies:

Firstly, to maintain a consistent level of user experience, sufficient capacity must be provisioned to link users with their applications whether they are local, remote or hosted on a cloud platform. This drives enterprises to model corporate infrastructure on Internet content delivery principles of network flattening and decentralization. In parallel, connections to third-party services, such as public cloud platforms, must be scaled up to allow seamless transition of workloads and data between private and public infrastructures.

As critical business sensitive information is travelling greater distances between facilities it may transit leased fibers, third-party circuits and public networks. High bandwidth and strong encryption is a key requirement to prevent exfiltration of confidential information.

The availability of feature rich 400GbE routers, combined with 400G-ZR and Arista's OSFP Line System (OSFP-LS) addresses the challenge of cost effectively constructing high bandwidth infrastructure for the enterprise backbone. Arista's portfolio of systems with line-rate MACsec encryption for speeds from 1G upwards provide strong encryption within the data center without impacting performance, while 400G MACsec and IPSec are ideally suited to provide encryption between facilities to ensure sensitive data is kept secure wherever it travels.

Consolidation

In the context of the 400GbE technology cycle, consolidation may be interpreted in two ways, both ultimately resulting in the ability to achieve more with less. For many organizations, current 100GbE links provide sufficient headroom for the mid-term, nonetheless 400GbE systems and interfaces can be considered as a viable alternative.

Two significant advantages of 400G devices are:

- Extremely high bandwidth densities (4x or greater than previous generation platforms)
- The highest degree of flexibility to support current and emerging link speeds, including 10,25,40,50,100,200 and 400GbE

For smaller networks that are not currently bandwidth constrained by 100GbE, 400G of bandwidth may be unnecessary, however, in addition to the ability to support a raft of higher speed options, 400GbE platforms provide the ability to deliver substantially more 100GbE ports in a smaller footprint through high density OSFP/QSFP-DD interfaces.

Each 400GbE port supports up to 8 parallel lanes (twice the number of QSFP100) with each lane being capable of 50Gbps, twice as fast as QSFP100. Together these 2x2 improvements result in a quadrupling of the effective port density per system.

Based on the previous generation technology many fixed data center class products offered up to 32 ports of 100GbE in a 1RU form factor, requiring either multiple devices or deployment of modular systems to increase capacity and naturally losing some percentage of ports to interconnections to other switches.

The equivalent 400GbE device offers four to eight times the density - up to 128 x 100GbE ports in a single rack unit:

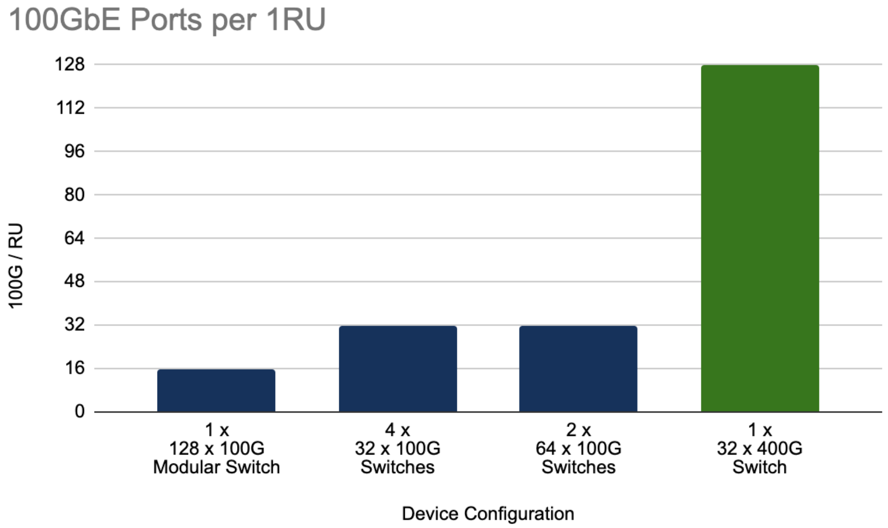


Figure 3. Relative density in 100Gbps/RU of switch generations

With up to 128 x 100GbE available in a single 1RU switch, networks can be greatly simplified, reducing both the device and cable count and flattening out multi-tiered networks resulting in a lower TCO and by removing oversubscription a higher performing, bottleneck free environment.

In addition, with flexible support for speeds from 10G-400G, modern 400GbE devices can cater for existing link speeds, while providing future proofing for 50, 200 and 400GbE connections as they become a requirement. Arista's 7050X4 and 7358X4 are the platforms of choice for enterprises investing in 50G, 200G and 400G for the data center.

Consolidation in HPC

For high performance networks and compute clusters, consolidation is also an important consideration. In general, large clusters benefit from deterministic performance between any two nodes in the network. Single or two tier networks are the optimal solution for providing deterministic, high-bandwidth interconnectivity, but are limited in scale by the capacity of the network spine devices. The larger the spine switch, the higher the potential radix of the cluster.

To illustrate this, an example compute cluster with single-homed 25GbE compute nodes connecting to a 4-way leaf-spine fabric with 3:1 oversubscription is shown below. In this model, each leaf switch supports 48 compute nodes and has 4 x 100GbE connections across the spine switches:

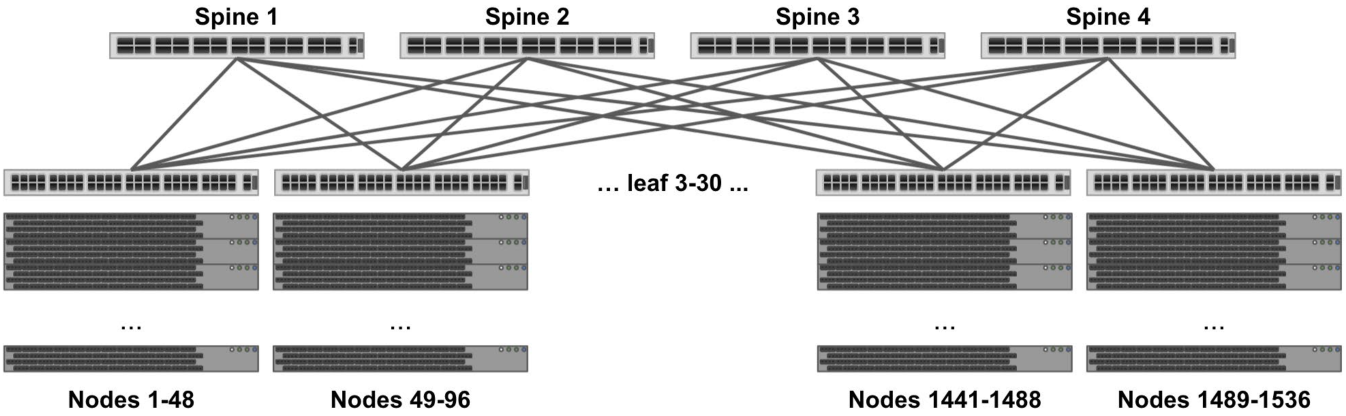


Figure 4. A typical 3:1 oversubscribed cluster with 25G attached nodes and 4-way 32x100G spine switches

With an architecture based on 32 x 100GbE 1RU spine switches, the maximum number of compute nodes that can be supported in a single cluster is 1536 servers across 32 leaf switches. Increasing scale beyond this number requires multiple clusters that are themselves interconnected via an additional spine layer, consuming some of the spine-leaf ports and resulting in a further level of oversubscription between nodes in adjacent clusters which results in non-deterministic edge-to-edge performance.

For example, the following diagram depicts deployment of a super-spine that allows 4 clusters to be interconnected at the cost of a further 3:1 oversubscription ratio between clusters. In this model, 8 x 100GbE ports from each cluster spine connect in pairs across the four Super-Spines resulting in a smaller number of nodes per cluster:

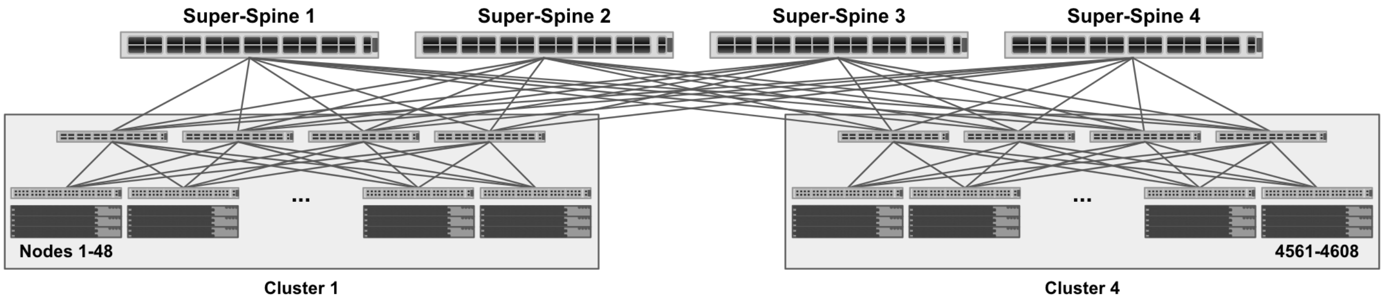


Figure 5. Scaling a 4-way cluster through addition of a super-spine tier

In this example, the addition of a super-spine layer provides only 3x the scaling of a single cluster to a maximum of 4608 compute nodes, despite adding considerable cost in switches and cabling as well as operational complexity. Application performance will also suffer between hosts located in different clusters due to a higher degree of oversubscription (9:1) and increased latency across two additional switch hops.

By contrast, leveraging a 32 x 400G device configured to offer 128 x 100G ports, a simple 2-tier fabric is able to support up to 6144 servers across 128 leaf switches with a fixed 3:1 oversubscription and optimally deterministic bisectional performance. The following diagram shows each leaf connecting with 4 x 100GbE across the four spine devices.

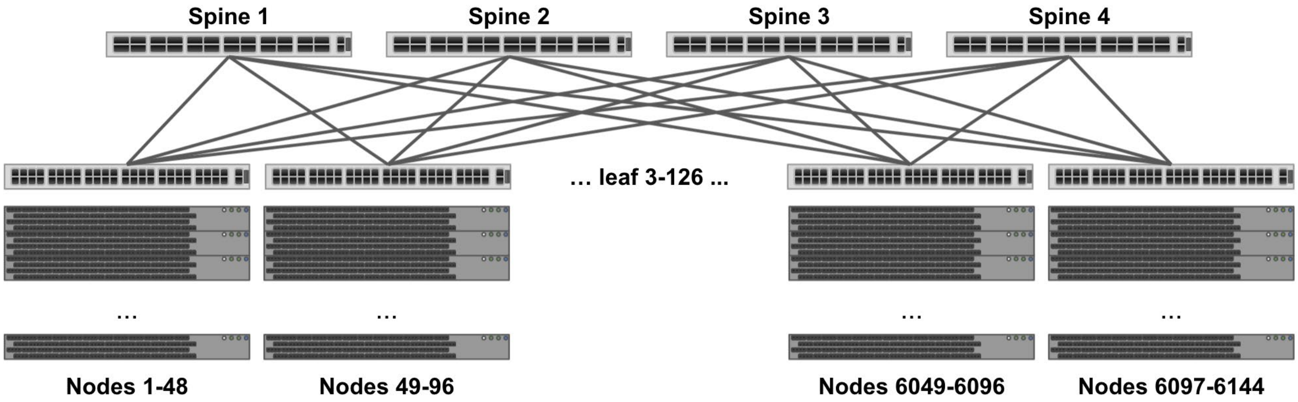


Figure 6. A typical 3:1 oversubscribed cluster with 25G attached nodes and 4-way 32x400G spine switches

The table below summarizes the options discussed above:

Table 1: Comparison of Cluster Configurations

Spine Device	# Spine Switches	# Leaf Switches	Total RU of Switches	Max Compute Nodes	Edge to Edge Oversubscription	Compute Nodes per RU of Network
32 x 100G	4 Spines	32	36	1536	3:1	42.6
64 x 100G	4 Spines	64	72	3072	3:1	42.6
32 x 100G Spine and Super-Spine	4 x 4-way clusters + 4 Super-Spines	96	116	4608	9:1	39.7
32 x 400G (128 x 100G)	4 Spines	128	132	6144	3:1	45.5

It is easy to see how increasing the density of the spine switch has a significant impact on cluster scalability, edge to edge performance and the ratio of compute nodes to network devices. Deployment of high density 400GbE devices such as the 7050X4 and 7060X4 platforms as 100G spine switches can enable consolidation of complex topologies into efficient, high radix leaf-spine models.

For yet higher scaling, the deployment of larger 400G devices, up to 64 x 400G, or the 7800R3 series of modular systems further enables clusters to scale to extremely large numbers while preserving two network tiers.

Cost Reduction and improved Power Efficiency

Consolidation itself drives cost reduction through simplification, reducing the number of devices deployed and in turn the overhead of operations and maintenance. For an organization considering a green-field deployment, the choice must be made between which generation of technology is most appropriate at each tier of the deployment.

When making the choice between the previous generation of 100G and newer 400G products there are several network roles where 400G systems should be considered as more cost efficient means to deliver high density 100GbE compared to the 100G focused alternatives.

400GbE systems benefit from significant improvements in silicon design and fabrication, moving from the 16nm process node to 7nm increases the transistor density by a factor of 3.3 and a power reduction of approximately 65% per transistor. These improvements in silicon density allow for higher speeds, higher density, lower power and more functionality in switch silicon architectures.

Reviewing a comparative cost per 100GbE port, it is clear that a fixed format 32 x 400G device offers a capital cost advantage over 100G focused platforms, especially where the opportunity exists to replace a multi-chip modular system.

Providing 128 x 100GbE from a 400G device can offer a cost per port saving of up to 75% over 100GbE modular technology, simultaneously reducing the physical footprint by over 87%:

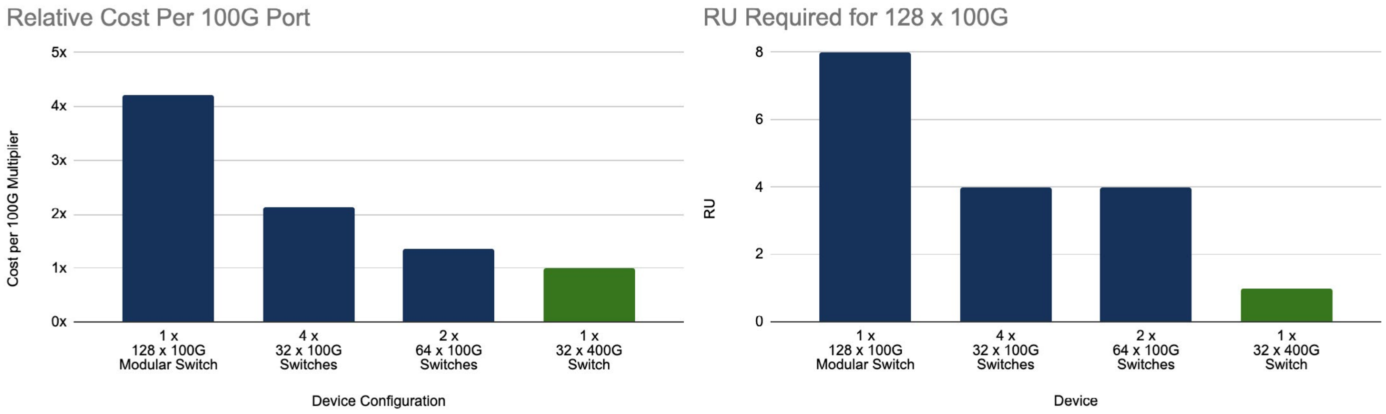


Figure 7/8. Comparison of the relative cost and physical space requirements per 100Gbps of bandwidth

Power savings are also significant, with a 45%-80% reduction in energy usage per 100GbE port to achieve the same density:

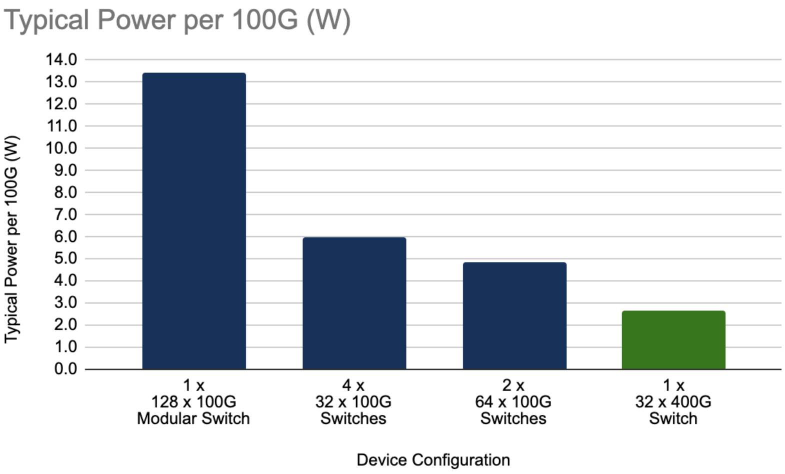


Figure 9. Comparison of the relative power consumption per 100Gbps of bandwidth

The combination of savings in initial purchase cost, space and power consumption create opportunities for long term cost savings and reduced environmental impact. Additionally, by reducing the ratio of infrastructure to compute, more power and space can be dedicated to application deployments, extending the capabilities of a facility.

400GbE systems also provide a level of future-proofing for predominantly 100G environments, with support for a broad range of current and future link speeds and access to advanced data center interconnect (DCI) technologies like 400G-ZR.

While first generation 400GbE technology commanded a premium and were suited to a few use cases, Arista's broad portfolio of 400GbE systems in both fixed and modular systems is available to all organizations. 400GbE systems can be considered for many traditional dense 100GbE roles and have a dramatic impact on costs and efficiency.

Summary

The 400 Gigabit Ethernet ecosystem delivers more than simply the next bandwidth increase, coinciding with multiple parallel technologies such as 400G-ZR, Arista's OSFP-LS, PCIe Gen4, SmartNICs and NVMe, in addition to application and workload trends including increased use of AI/ML and cloud native applications to offer a robust solution for emerging workloads and the network topologies optimized for them.

Even where 400GbE itself is not an immediate need, the flexibility of new platforms combined with significant cost, energy and footprint reductions means they should be considered for both brownfield consolidation and greenfield projects as a viable alternative to legacy 100GbE systems.

Arista's portfolio provides 400GbE technology solutions for data centers, DCI and backbone networks, both for organizations with extreme workloads as well as delivering the core benefits of this next-generation technology to a broad range of use cases that are not bandwidth constrained today. The range of platforms include the X-Series and the R-Series range of fixed and modular systems built on the rich functionality of the common Extensible Operating System (EOS) operating system, minimizing the learning curve to adopt next generation technologies.

Santa Clara—Corporate Headquarters
5453 Great America Parkway,
Santa Clara, CA 95054

Phone: +1-408-547-5500
Fax: +1-408-538-8920
Email: info@arista.com

Ireland—International Headquarters
3130 Atlantic Avenue
Westpark Business Campus
Shannon, Co. Clare
Ireland

Vancouver—R&D Office
9200 Glenlyon Pkwy, Unit 300
Burnaby, British Columbia
Canada V5J 5J8

San Francisco—R&D and Sales Office
1390 Market Street, Suite 800
San Francisco, CA 94102

India—R&D Office
Global Tech Park, Tower A, 11th Floor
Marathahalli Outer Ring Road
Devarabeesanahalli Village, Varthur Hobli
Bangalore, India 560103

Singapore—APAC Administrative Office
9 Temasek Boulevard
#29-01, Suntec Tower Two
Singapore 038989

Nashua—R&D Office
10 Tara Boulevard
Nashua, NH 03062