



WHITEPAPER

The Rack Is The New Server

The Register®



The supercomputer industry and the web infrastructure industry – what are sometimes called hyperscalers because they need millions of servers support billions of users on their applications – have much in common. Rackscale infrastructure is one of them.

And soon, this rackscale approach to building and deploying datacenter equipment will be common not only among HPC centers, hyperscalers and their cloud computing groups, and other large infrastructure service providers, but among enterprises who are mainstreaming generative AI to extend their applications.

As is often the case with technology, those who are doing high performance computing – pushing the limits of compute and storage scale – blaze the trails for fast-following enterprises. Supercomputing centers and hyperscalers both use distributed computing systems for a handful of applications that need to be run at a scale that is not typical in a normal enterprise. Both also have strong traditions of designing and building their own systems, but it has been a long time since either actually cobbled together servers from parts and then networked the resulting servers into a distributed system. What was possible with dozens to

hundreds of servers is no longer possible with thousands to tens of thousands of servers.

And so the hyperscalers and HPC centers of the world have increasingly relied upon original equipment manufacturers (OEMs) or original design manufacturers (ODMs) to turn the compute, storage, and networking specifications that these customers come up with into specific server and cluster designs, and as the years have gone by, they have also increasingly relied upon them to provide complete racks of servers, storage, networking, power distribution, and now liquid cooling that roll off the trucks straight into the datacenter and only require two or three cables (depending on the design) to be attached to them so they can start running workloads. The first one is power, the second one is networking, and the third one is liquid cooling running out to a chiller.



ART: Supermicro SuperCluster with Liquid Cooling (512 GPUs in 8 racks)

In effect, the rack is the new server, the datacenter hall is the new rack, and the datacenter (a collection of interconnect halls) is the new row, and a collection of datacenters commonly called a region is a new beast in its own right, aggregating hundreds of thousands to millions of servers (some doing compute, some doing storage) into a single fabric that can be made to look like one giant system if need be or carved up into pieces on the fly to be used by many customers at once.

Just as the level of abstraction for systems has broadened, so have the mandates of the companies that we still call server manufacturers. When it comes to generative AI training and inference at scale, traditional high performance computing simulation and modeling, and advanced data analytics and risk management systems common in the enterprise, these companies are more accurately described as turnkey supercomputer makers. Supercomputing has finally gone mainstream, and these OEMs and ODMs are putting together rackscale architectures that allows them to build these quickly and at scale – requirements for winning customers and keeping them.

Pieces Of Eight

In modern rackscale supercomputers, eight is the magic number. There are eight GPU accelerators on an NVIDIA HGX™ platform and that means there needs to be eight SmartNICs or DPUs on the server node housing those GPUs so there is a one-to-one pairing between network ports and GPUs so any GPU can talk to any other GPU in the distributed system – or all of them at the same time for certain all-to-all or all-to-one collective operations that are common in both HPC and AI workloads.

On the host machine, there is typically a single socket or dual-socket X86 server node that acts as an application controller and serial processor for the AI or HPC software stack. But if you look carefully at the architectures there are usually eight chipllets in a socket or eight NUMA regions on a pair of processors paired with these eight GPUs on the system board. So, for instance, on the current “Hopper” H100 generation of HGX systems, there is a pair of X86 processors that have four chipllets each, for a total of eight chipllets that correspond to eight NUMA regions that can be

isolated and paired to each GPU. On the nodes used in the “Frontier” supercomputer at Oak Ridge National Laboratory, the custom X86 processor in the host has eight chipllets, and the node has four dual-chip GPU accelerators – again a one to one pairing of eight GPUs.

The rackscale infrastructure scales up from there. For instance, four of these HGX systems can be put into a single rack, for a total of 32 GPUs, and either be air cooled or liquid cooled as the environmental of the datacenter allow. A row of these machines has a rack of networking in the middle and four racks of CPU-GPU nodes on either side for a total of eight racks of compute. To scale this further, you add rows until you fill the data hall and you add data halls until you fill the datacenter. If you can get allocations for enough GPUs to fill a data hall or a datacenter, you are doing a lot better than most organizations and you are probably training pretty large AI models.

And, you are probably buying rackscale infrastructure rather than buying servers, switches, cables, racks, and power and cooling distribution units and cobbling this all together yourself. Time is money, and time not using a GPU while you are setting up an AI supercomputer turns out to be worth a lot of money. Moreover, due to demand for GPU accelerators being several factors higher than supply, GPU vendors are both being very selective about who they give GPU allocations to and how many they give as well, and the primary gating factor being used by these GPU makers after they look at the price paid is how quickly customers can get machinery in the field and start getting real work done with the GPUs. Fortune favors the fast.

And hence, given the grief of building clusters with hundreds to thousands of nodes and the time it would take to do so, rackscale is the new norm for setting up these complex AI supercomputers.

Rack ‘em and stack ‘em

So why rackscale, and why now? It is not like rackscale architecture is a new idea in the datacenter.

For instance, Egenera, with its BladeFrame rackscale designs for integrated server, storage, and switching, got its first

product in the field in 2001. Egenera had trouble getting customers to adopt its technically elegant designs – mainly because server virtualization had not yet gone mainstream on X86 systems and compute, memory, and storage were still very tightly coupled and could not scale independently of each other.

Rackable Systems was founded in 1999 as the Dot-Com Boom was preparing to go bust, but made it through that downturn and within a few years had Microsoft, Yahoo, and Oracle as customers for its rackscale machines – at numbers that sounded like a large amount of infrastructure at the time but was small potatoes by today’s hyperscale standards.

While many hyperscalers were interested in the Intel ideas, Meta Platforms – which was just Facebook at the time – had launched the Open Compute Project two years earlier and was getting traction with its own Open Rack designs. Facebook had lined up a number of manufacturers who were interested in building to its rackscale specifications as well as dozens of customers who wanted to deploy the same infrastructure that Facebook had designed and use the manufacturing supply chain it was trying to foster. To a certain extent, the Open Rack designs from Meta Platforms as well as the Microsoft Open Cloud Server that dates from the same time and Project Olympus rack design that was unveiled in 2014 (both donated to OCP) have been successful in promulgating the idea of thinking about infrastructure at the rack level rather than the server level.

Right on the heels of the Open Compute Project launch by Facebook, the Open Data Center Committee, also known as Project Scorpio, was launched by Baidu, Alibaba, and Tencent to come up with a shared rackscale design to leverage their combined scales, and two years later, Intel launched its Rack Scale Architecture to try to create standards for packaging, disaggregation, and composability inside and across racks, including its own optical interconnects. Intel has largely folded its rackscale efforts into the OCP and the ODCC projects, which accounted for 7 percent of worldwide server shipments in 2016 but which are projected to grow to around 40 percent of server shipments by 2025. (Thanks in large part to the rackscale designs offered up as standards by Meta Platforms and Microsoft.)

The rackscale concept just got a boost from NVIDIA with its latest “Blackwell” architecture announcements. NVIDIA has been architecting rackscale systems, which it calls SuperPODs, for the past four years. Starting with its “Ampere” architectures, NVIDIA HGX A100 GPU compute complexes, and 200 Gb/sec HDR InfiniBand networking. But architecting such SuperPODs and deploying systems based on them for its own internal use is not the same thing as being a volume manufacturer of them. With the Blackwell-based DGX GB200 NVL72 system that will be shipping in late 2024, NVIDIA has created a rackscale system with 72 GPUs linked together with its own NVIDIA® NVLink® Switch 4 fabric and to a total of 32 of its “Grace” Arm server CPUs. With this design, the rack is literally the new node from an architectural perspective. But NVIDIA is still not going to be a volume manufacturer of these very innovative Grace-Blackwell systems.

More importantly, NVIDIA only builds platforms from the components that it sells. But a lot of organizations want – and need – a choice for every component in the stack, but they want to use the same architectural process and manufacturing partners to create different kinds of clustered systems. This is how and why Supermicro has created its SuperCluster rackscale systems.

Time Is Always Money, Too

And of course, you don’t have to adhere to either OCP or ODCC standards, or be Meta Platforms, Microsoft, NVIDIA, or Intel, to create rackscale solutions. What you need, really, is a factory and control of design of system components and rack components. And then customers no longer can buy servers, switches, and storage and integrate them inside their own facilities can offload that job to companies that not only have manufacturing skills but who can also solve particular architectural problems, offer a choice of componentry for building rackscale systems, and add complimentary services on top of the buildout of these systems.

As it turns out, the number of such customers who need someone to build rackscale systems and cluster many of them together is growing just as the capacity of Supermicro, a handful of ODMs such as Foxconn, Inventec, Quanta Computer, Wistron, and a handful of OEMs like Hewlett

Packard Enterprise, Dell Technologies, Lenovo, and Inspur is also growing to meet that demand. Some are growing faster than others, but none are growing as fast as Supermicro.

Companies have better things to do than be their own system integrators these days, and with the expensive and complexity of AI training and inference systems in particular, they simply do not have time to waste. If it takes three to six months to cable together an AI system and test it, and that machine costs \$500 million to \$1 billion, that time that the machine is not being used is very expensive indeed if the machine is only expected to be in the field for three, four, or five years. That is somewhere between \$50 million and \$83 million worth of lost value for a \$1 billion machine if it takes three months to build it. And it is also three months lost on doing AI training runs that typically take that long to complete. So there is the cost of lost opportunity for any delays in deployment, too.

Supermicro is uniquely positioned as a provider of rackscale systems and full HPC and AI clusters in a number of ways, which gives it advantages over other ODMs and OEMs. And this is why Supermicro's rackscale business has been skyrocketing in the past two years and why it is expected to grow its overall systems business to \$25 billion a year and beyond over the next several years, rivaling HPE and Dell as high volume system makers.

First, Supermicro has a long history of engineering system components, from motherboards and peripheral cards to system enclosures. Supermicro also knows how to engineer storage servers and switches. And over the years, as companies have decided that they don't want to deal with unboxing all of their new gear, putting it into racks, networking it together, installing software, and testing the whole shebang, Supermicro has evolved from a maker of motherboards, peripherals, and enclosures, to a high volume system and storage maker, and has grown into one of the highest volume shippers of rackscale systems in the world.

Rackscale Manufacturing Is Harder Than It Looks

Two things are immediately obvious. If you are going to buy at rackscale, then you need a partner who can build at

rackscale. And if you need to build and test at rackscale, your factory has to be a special kind of datacenter. And increasingly, as companies decide to just let go of the whole system integration headache, that factory has to have software experts and tools to load up a complete software stack – for servers, storage, and switches – on the cluster of rackscale machines and test it to make sure it all works together as expected. Customers want to buy a total solution, and NVIDIA wants for its partners to sell a total solution as well, including testing and validation of hardware and software such as the NVIDIA AI Enterprise stack at the cluster level.

This is not your father's server business. This is actually the supercomputer business as it has operated for decades going mainstream to hyperscalers, cloud builders, service providers, and large enterprises the world over.








In system manufacturing, there is a nomenclature that is used to describe the kind of manufacturing that is done.

With Level 10, the manufacturer assembles a single server, and if you want, you can buy multiples of these and do rack-level integration yourself. This is what the OEMs have done for years, and it was quite sufficient for a long time given the relatively modest computing needs of most enterprises.

The hyperscalers and cloud builders had deployment scale and system configuration needs that were outside this L10 experience, and so the ODMs started co-designing custom hardware with these customers and delivering them in rackscale modules, complete with servers, switching linking them together, power distribution units, and storage if necessary. This level of manufacturing, called L11, is all about very high volumes, and it only involves hardware and its firmware.

Going up another notch is L12, short for Level 12, and it is here that manufacturers build at the cluster level, stitching together different kinds of rackscale systems together – infrastructure servers with just CPUs, AI and HPC servers with GPUs and other kinds of accelerators, leaf and spine switches that implement the network topology for the cluster, and add in the operating systems, file systems, and middleware from the customer that make it a running system.

Delivering Plug-and-Play Solutions at Rack Scale

	Application Ready	Cloud, HPC, AI/ML, Data Analytics, ERP, SQL, 5G, IoT
	Cloud Ready	OpenStack, OpenShift, RHHI, Azure Stack HCI, VMware vSAN
	Cluster Ready	NVIDIA AI Enterprise Software, SuperCloud Composer, Ansible, BCM, and etc.
	File System Ready	Ceph, WEKA, VAST, Scality, OSNexus and etc.
	OS Ready	RHEL, SUSE, Ubuntu, CentOS, Windows
	Network Ready	10G/25G/40G/100G/400G Ethernet, Infiniband EDR/HDR, Omni-Path
	Hardware Ready	Systems optimized for NVIDIA HGX, MGX and PCIe GPUs, Rackmount, Twin/Multi-node, Blade, and storage servers, Liquid Cooling

ART: Supermicro plug and play stack

Increasingly, the systems with GPU and other kinds of accelerators are also liquid cooled to improve their energy efficiency and their density, and the L12 integration includes rear door rack coolers or direct liquid cooling, or a mix of the two approaches.

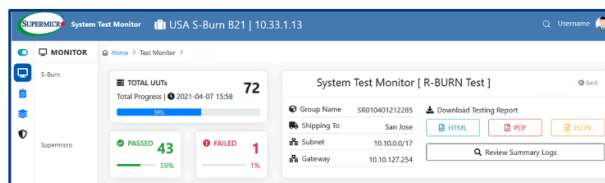
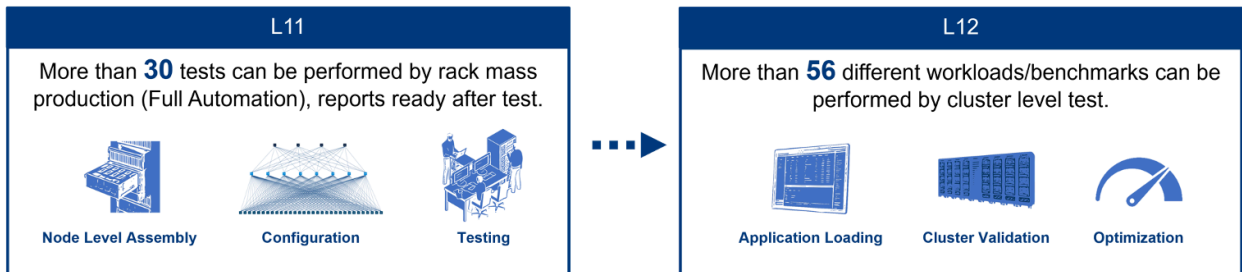
In the end, with an L12 manufacturing partner, all customers have to do when they receive their shiny new machinery is plug in those two or three connections mentioned above and the rack is up and running and ready to take on applications.

Aside from the shorter time to application, there are other benefits to L12 manufacturing.

First, the manufacturing can be more predictable. These days, allocations of key compute engines, network interface or switch devices, and other components are often based on the ability to get up and running fast. So if a GPU vendor doesn't think you can get up and running quickly, you loose out on device allocations.

Supermicro L11/ L12 Solution Validation Lab

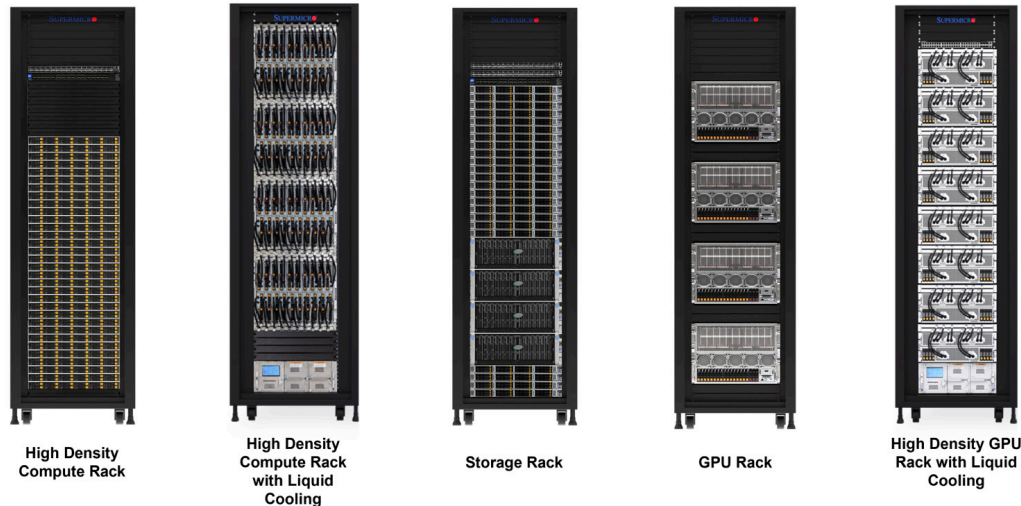
L11 and L12 with flexibility to accommodate customers' specific testing requirements.



GPU Test and Validation Examples: NVIDIA AI Enterprise Software, NVQual, cuda_mem, GPU validation, GPU linpack, HPGC, NCCL, RCCL, GEMM, Bandwidth, GPU MemoryPerf, DeviceQuery, MultiCopy, KFDTest, RVS, TransferBench, GST and gpu_burn, ResNet-50/101 and ImageNet etc.

ART: Supermicro l11 l12 manufacturing

Liquid-Cooled Rack: Sample Configurations



ART: Supermicro liquid-cooled rack design

Another benefit is that the consistency in the architecture of rackscale devices speeds the manufacturing process – the people who build one rack build the next two, the next dozen, the next hundred, the next thousand.

The standardization of rackscale designs also drives down errors. This is why you find the hyperscalers and cloud builders only typically have a half dozen rack designs. This drives down costs through volumes and drives up efficiencies. At the L12 level, where the hardware is tested as a full cluster, you get the added benefit of knowing that the machinery is immediately available as it rolls into the datacenter.

While there are significant savings that come from buying at rackscale and at clusterscale, for most customers the time to value is far more important than any potential savings from buying a pre-integrated cluster.

Liquid Cooling Will Be Normal For Most AI And HPC Systems

Right now, somewhere between 5 percent and 10 percent of the systems that are installed by Supermicro have liquid cooling of some fashion, but the company believes that by 2025, that share of systems is expected to rise to around 20 percent. The reason is that the most powerful accelerators

that will be used to run AI and HPC workloads are pushing up above 1,000 watts in power draw.

Supermicro has the ability to integrate with any liquid cooling technologies that customers may desire in their rackscale designs, but because standardization is lacking when it comes to liquid cooling, Supermicro designs and builds its own coolant distribution units (CDUs), cold plates for compute engines and memory, and cooling manifolds as well as its own cooling towers, which come in three sizes.

Add it all up and Supermicro can power and cool racks with up to 80 kilowatts of equipment, and will soon be able to do 100 kilowatts per rack, and has the ability to push that up to 250 kilowatts per rack when it becomes necessary to do so.

All of this capability is great, but customers buying at rackscale are building larger and larger clusters, too. So the manufacturing scale matters as much as the ultimate system scale. In 2022, Supermicro could build around 2,000 racks per month, and here in early 2024, it can do 4,000 racks per month and by the end up June it will be up to 5,000 racks per month, with about 1,500 racks per month of liquid cooled gear.

It will be hard to find a manufacturer that has facilities in the United States, in Europe, and in Asia – and thus can address the concerns of sovereign manufacturing for the major regions of the world – and has them at the scale of Supermicro.



ABOUT SUPERMICRO

Supermicro is a global technology leader committed to delivering first-to-market innovation for Enterprise, Cloud, AI, Metaverse, and 5G Telco/Edge IT Infrastructure. We are a Rack-Scale Total IT Solutions provider that designs and builds an environmentally friendly and energy-saving portfolio of servers, storage systems, switches, software, along with global support services.