WHITE PAPER

10 BEST PRACTICES FOR CSP'S TO SCALE THE DATA CENTER

Considerations To Create a Modern Data Center

TABLE OF CONTENTS

Executive Summary

Service providers face a myriad of challenges in today's dynamic environment, including Cloud Service Providers (CSPs), Managed Service Providers (MSPs), Software-as-a-Service (SaaS) providers, and enterprise private cloud operators. The landscape of technologies that constitute a modern data center is evolving swiftly, with cost management being a perennial concern. Moreover, the expansion of services does not always correspond with an increase in staffing. Partnering with a reliable supplier that provides cutting-edge servers, storage, and networking solutions, pre-tested and assembled into a rack and pre-configured with the necessary software stack, can alleviate some of these challenges, leading to quicker deployment of new services or enhancement of existing ones.

As a leader in supplying rack-scale solutions to CSPs, both large and small, Supermicro has significant experience in product development but also supply chain logistics, service and support, and sizing and testing.

Below are ten best practices for architecting and deploying many rack-scale data centers based on real-world experiences working with customers worldwide.

#1 Standardize and Scale

Many hyperscalers have adopted a "standardize and scale" hardware deployment model. In this strategy, a standard configuration of compute, storage, and networking is selected, and then this configuration is scaled up and/or down so that there are multiple configurations. These can be considered small, medium, and large configurations that can be deployed at various data center sizes depending on the number of simultaneous users, workload sizes, and growth estimates.

#2 Optimize the Configuration

This advice contradicts advice #1, which is to standardize on specific configurations at the server and rack level. Workloads can vary significantly, depending on the CSP offerings. If a CSP specializes in a particular workload, using a repeatable rack-scale infrastructure will work; however, in many cases, a more varied workload will be the norm.

At the start of the process, running the actual or representative portion of the software on the rack configuration is essential to determine the best configuration of CPUs, including the number of cores, memory, storage, and I/O. For example, the average memory cost is 49% of the total server cost, according to the Semiconductor Research Consortium (SRC). If a rack is dedicated to a particular application that utilizes a smaller memory footprint, and the SLAs can remain, this can substantially reduce the cost of the servers. Similarly, choosing the suitable CPU cores and clock speeds for the application will also lower the TCO while maintaining the required performance.

GENESIS CLOUD

"Genesis Cloud, using the most powerful Supermicro AI servers available, continues to offer customers the most advanced AI Training servers available today. Our data centers are powered by green energy, allowing users fast, low latency access to the Supermicro GPU servers with the latest NVIDIA HGX H100 8-GPU hardware. We continue to work closely with Supermicro to offer our customers the most advanced GPU servers in the industry".

Dr. Stefan Schiefer, CEO, of Genesis Cloud

In a public and shared cloud, there may be various workloads, depending on the specialty of the CSP. Different sets of optimized configurations can be set up, with the applications sent to the other groups of servers. For example, AI training workloads would be sent to the GPU optimized servers. In contrast, a database application would be sent to and executed on many standard two socket systems without GPUs.

So, a trade-off between standardizing configurations and optimizing for specific workloads must be considered.

NETFIRE

"NetFire can respond to a wide range of customer requirements using our Supermicro AMD based servers. We can offer customers a range of implementation environments, from using our public cloud to on-premises clouds. The performance per watt of AMD CPUs, in different Supermicro servers, helps customers to reduce costs while maintaining high performance responses, supporting increased workloads when required."

Bart Matusiak, President and Founder at NetFire

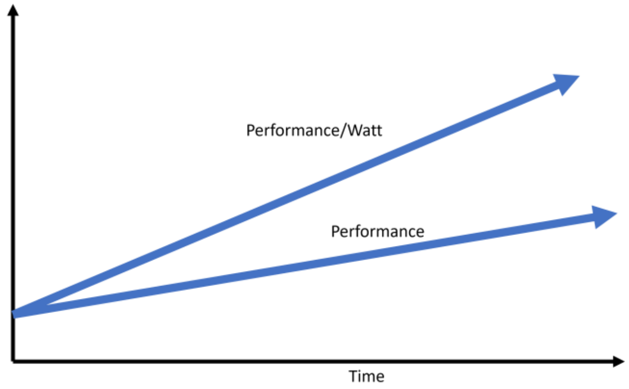
#3 Plan for Technology Refreshes

Change is a constant in technology. However, waiting for the latest and greatest technology is futile as new technologies and improvements are invented. Strategically planning around critical technology transitions and implementing an upgrade and/or migration strategy can maximize the benefits to the buyer. A supplier with deep partnerships with key technology suppliers who can share the transition plans, cost impacts, and supply chain issues with you is critical. In addition, a disaggregated or modular server and rack approach can mean upgrading specific components or servers without replacing the entire chassis. New generations of servers that can perform substantially more work per watt may also need more power. The design of a new data center should not be limited by rack power requirements when the initial servers and racks are installed. By working closely with your supplier, understand the criteria for potential technology in the data center.

ABSOLUTE HOSTING

"We launched our VPS server products by advertising on social media and local IT forums. Within a couple of months, we'd reached our first 100 VPS servers. We ran a public beta for potential clients and anybody within the local IT community to stress test. The feedback we received was phenomenal. Everybody was blown away by how powerful these AMD EPYC CPUs were. This enthusiastic feedback helped to solidify Absolute Hosting's reputation as a leading provider of reliable and high-performance VPS services."

Jade Benson, Managing Director of Absolute Hosting



#4 Look at New Architectural Approaches

New technologies can increase performance at lower costs. For example, depending on the required SLAs, code base, and matrix processing level, AI workloads can be done on CPUs or GPUs. Some workloads can be moved from the CPU to an auxiliary DPU (Data Processing Unit), which acts as both a network interface and a data processing unit. Some workloads benefit from a custom approach using an FPGA. The introduction of CXL 2.0 (Compute Express Link) provides another layer in the memory hierarchy below directly attached to DRAM but above SSDs. For example, this enables the concept of pooled memory, which can be flexibly allocated to one of the CPUs on the system, and mitigates the issue of stranded memory, which is directly attached to a CPU but not fully utilized. These new technologies may benefit the particular workload and software stack for the intended service. Testing new technologies in a Proof-of-Concept (POC) setting before large-scale deployment is essential. Working with a hardware partner on early POC testing with these new technologies is key to gaining advantages before your competitors.

ATLANTIC.NET

"Supermicro continues to provide Atlantic.Net with an excellent range of optimized servers. We are able to easily create innovative, highly performant, and dependable solutions that allow our customers to get the business results that they need. Supermicro's service and knowledge of multiple industries and workloads allow us to work together to bring advanced solutions to customers across multiple industries with varying resource and compute models."

Josh Simon, Vice President of Cloud Services and R&D at Atlantic.Net

#5 Have a Support Plan

Many businesses run 24x7, so the systems must also run to support these workloads. Working with a systems provider experienced with CSPs provides the flexibility of matching your support needs to their infrastructure, whether 4 hours and 24/7 or supporting a parts depot for self-maintenance. For support, determine what can be supported by the in-house team and outsource the rest to your provider. It's also crucial to ensure that your provider has tested your software stack to ensure hardware vs. software issues can quickly be escalated to the proper responsible party without delays.

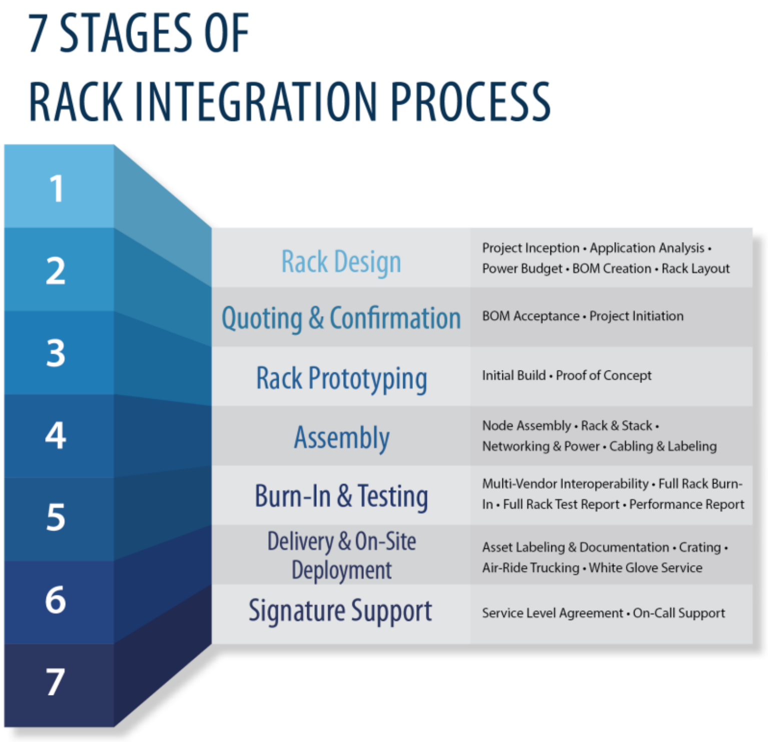
SEEWEB

"After evaluating various suppliers for our new flagship Cloud Service dedicated to AI/ML/DL, Seeweb decided to purchase Supermicro GPU servers that respond to all our needs. The GPU density of the Supermicro systems gives us the required performance, flexibility, and lower energy consumption that helps Seeweb to design the new Cloud Server GPU service with the right market fit. We are extremely satisfied with both the performance of the Supermicro GPU servers and our relationship with Supermicro" – Antonio Baldassarra, CEO at Seeweb

Dr. Stefan Schiefer, CEO, of Genesis Cloud

#6 Design for the Data Center

While the initial conversation may be about which server or servers to acquire for the desired workloads, the conversation will quickly turn to rack scale integration. As the number of racks at a site increases, it is necessary to understand the workings and limitations of the entire data center. The entire data center must be considered a whole unit, from topics such as the separation of cold and hot aisles, forced air cooling, and the size of chillers and fans to electrical distribution. The discussion of cooling technologies must be considered at the start since the data center's physical infrastructure will be different for air or liquid cooling.



7 STAGES OF
RACK INTEGRATION PROCESS

1
2
3
4
5
6
7

Rack Design	Project Inception • Application Analysis • Power Budget • BOM Creation • Rack Layout
Quoting & Confirmation	BOM Acceptance • Project Initiation
Rack Prototyping	Initial Build • Proof of Concept
Assembly	Node Assembly • Rack & Stack • Networking & Power • Cabling & Labeling
Burn-In & Testing	Multi-Vendor Interoperability • Full Rack Burn-In • Full Rack Test Report • Performance Report
Delivery & On-Site Deployment	Asset Labeling & Documentation • Crating • Air-Ride Trucking • White Glove Service
Signature Support	Service Level Agreement • On-Call Support

#7 Understand and Consider Liquid Cooling

New servers containing the latest CPUs and GPUs are quickly approaching the limits of air cooling, which requires a new approach, liquid cooling, to keep the microprocessors and accelerators running within their design limits. In addition, if the data center power budget is an issue, consider using liquid cooling to reduce the overall data center PUE (Power Usage Effectiveness) and minimize the HVAC cooling power. Many data centers have a 10-12 kW/rack budget, which becomes challenging for a full rack of servers, GPU servers, and storage. New systems for AI may each draw up to 10kW per server, resulting in an increased power per rack of up to 50kW. A liquid cooling solution allows for higher-density servers and GPU accelerated servers; the external heat exchanger is much more efficient than conventional HVAC cooling. A liquid cooling infrastructure must be planned in advance of the rack delivery time. Working with a company experienced in liquid cooling at the rack level is critical to an efficient data center.

ELIOVP

"Our relationship with Supermicro and AMD is extraordinary. We are extremely pleased with the responsiveness of both companies whenever an issue arises. The servers' performance is amazing, which increases our business, and reduces costs. By working with Supermicro, we can get new generations of servers with AMD technology earlier in our development cycle, enabling us to bring our products to market faster."

- Elio Van Puyvelde, CEO of Eliovp

#8 Measure What Matters

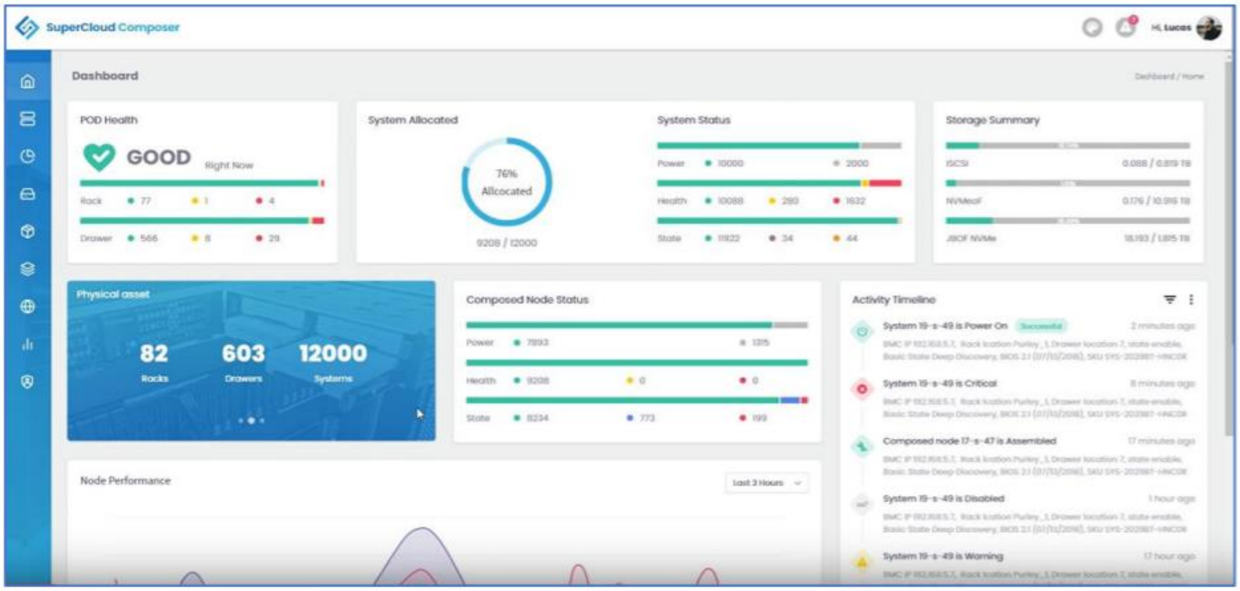
To assess the efficiency of your current data center, use instrumentation to measure CPU, storage, and network utilization. There are tools available to do this at the cluster level. These tools can provide valuable information on where existing bottlenecks occur and where over or under-utilization situations are not optimal. In addition, the temperature of the CPUs and servers can be measured, which can identify potential issues before the problems cause failures.

#9 Job Management

A data center for a cloud provider will most likely be used by many customers simultaneously, and therefore, a job management scheduler will be needed. With finite resources, not all requests for compute, storage, or networking may be satisfied, and jobs or applications will have to be scheduled or fit in as resources become available or until additional software can be acquired.

GLESYS

“GleSYS is very excited to work with Supermicro on current and future generations of servers. Our experience with Supermicro has been extraordinary. Not only did the systems work right out of the box, but we were also given direct access to program managers and service experts when needed.” - Jonas Törning, Purchasing Manager, GleSYS



INMOTION HOSTING

"Automating deployments of OpenStack and Ceph private clouds in under an hour at a low-cost entry point was a huge challenge. Supermicro embraced our vision and actively engaged their Networking, Blade, and OpenStack teams to help us bring this vision to market."

–Todd Robinson, President and Co-Founder of InMotion Hosting

#10 Simplify Your Supply Chain

It's often said that having "one throat to choke" in managing suppliers is optimal. While we're not advocating the threat of violence, simplifying the supply chain for key suppliers is a good idea for ordering, installation, and support. A single supplier who can provide servers, storage, networking, third-party software solutions, rack integration, and can integrate unique third-party hardware into a single system is optimal.

Now, two bonus recommendations:

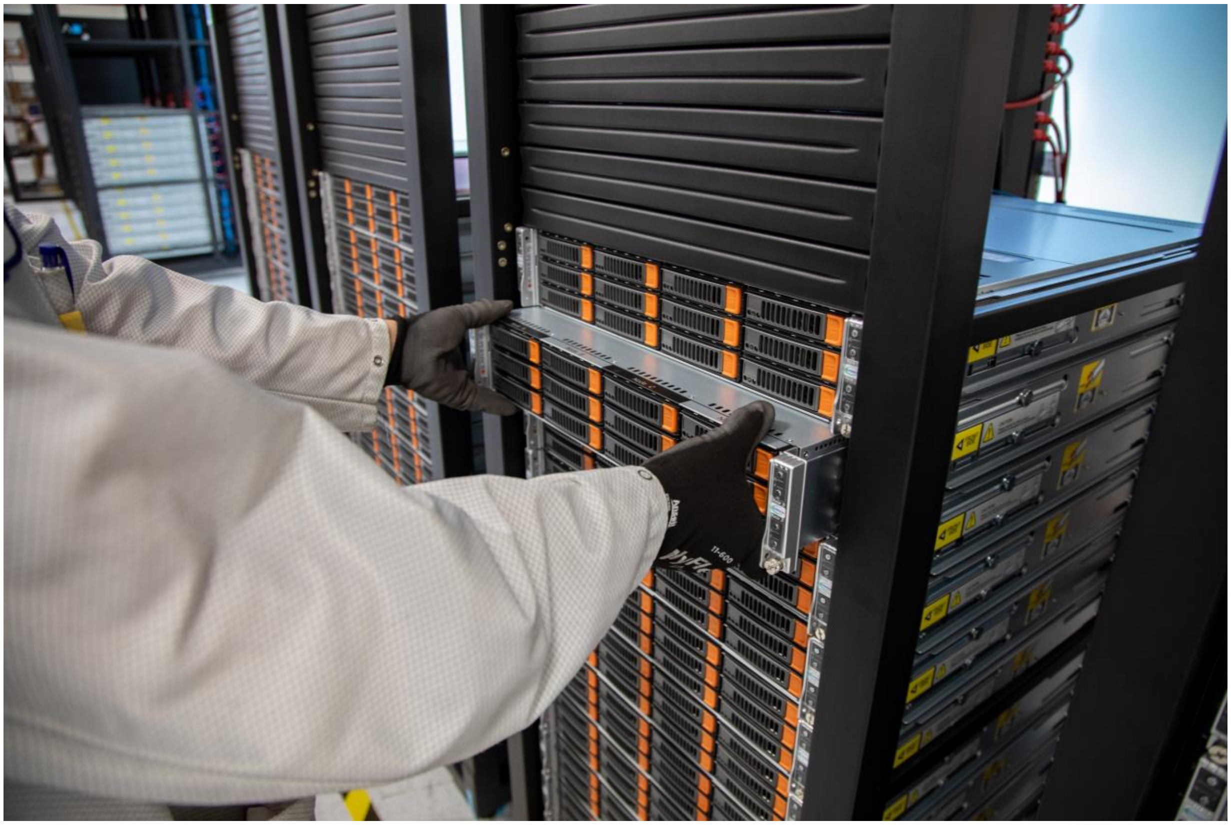
#11 Manufacturing Expertise Matters

It's an open industry secret that almost all of the large OEMs have outsourced their products' manufacturing, design, and supply chain to Original Design Manufacturers (ODMs) and Contact Manufacturers (CMs). The OEMs are mainly focused on marketing and selling these products. It is valuable to work with a company that designs all of its products, from chassis to motherboards and power supplies and manufactures these products in locations close to customers' locations. To the customer, this means that a data center supplier can be much more flexible, provide faster time-to-delivery, and ultimately reduce the TCO (total cost of ownership) through fewer intermediaries, faster transportation, and economies of scale.

NITRADO

"We appreciate the excellent relationship between our two companies, which enables us to use early ship programs with the availability of the latest hardware. Our customers and therefore also service quality has the highest priority for us."

Marcel Bößendörfer – CEO



#12 Experience Matters, Too

Putting all of your eggs into one supplier basket can be risky. So is adopting new technologies in the data center. Selecting a data center solution provider is no place for on-the-job learning or working with a company more focused on its own managed service offerings or making laptops. Working with a B2B company focused only on the data center and has been working for decades with service providers, large-scale HPC supercomputers, and powering solutions for the largest hyperscalers, OEMs, and enterprises.

VEXXHOST

"We are extremely excited and looking forward to using the new 3rd Gen AMD EPYC processors, especially the fact that they utilize the same sockets as our existing systems, and we will easily be able to take advantage of the new performance gains."

-Mohammed Naser, CEO at VEXXHOST

Summary

Planning and operating a data center as a service provider requires careful planning and a close working relationship with full service providers. There are a number of decisions to make that will affect the start-up times, SLAs, and efficiency of the data center. Whether designing and implementing a public shared data center or an on-premises data center, plan carefully, understand server and rack technology, and explore new technologies that will keep the data center running for years to come.

More Information

For more information, please visit: www.supermicro.com and www.supermicro.com/solutions/csp

SUPERMICRO

As a global leader in high performance, high efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs, and enables us to build and deliver application-optimized solutions based upon your requirements.