Data Center Refresh and Expansion Considerations

Keeping Up With Workloads

As organizations across the globe continue with their digital transformations, IT departments are being forced to deliver more performance and capability with static budgets. In addition, new technologies are constantly becoming available that allow for increased employee interaction and customer satisfaction. The workloads that IT organizations must support include new analytics platforms, artificial intelligence, machine learning, high-performance computing, cryptocurrency, virtual desktop interfaces, database access, and many more.

CONSIDER:

What are my most important workloads?

Keeping Up With Workloads

While there may be similarities across different companies' workloads, each company will have different requirements where application-optimized servers and storage systems are the keys to high performance at reduced costs.

Planning a hardware refresh requires an understanding of both corporate goals and IT requirements.
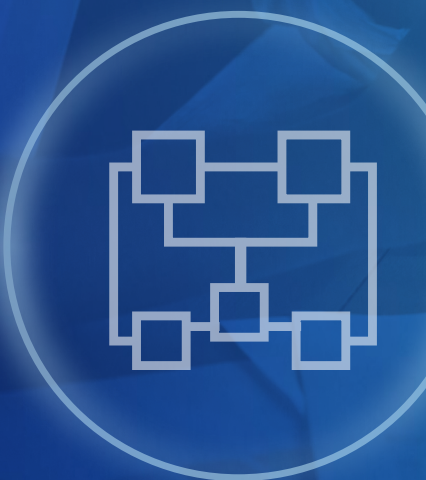
CONSIDER:

What are my most demanding workloads?

Why Refresh or Expand?

Understanding how a current data center performs is the first step to determining if a refresh is needed. An audit may be necessary when determining a refresh or expansion of a data center.

- **Are the applications delivering the performance expected?**

- **Looking forward, what will be the required delivery time frames, and for what types of applications?**

- **What percent of servers are considered "zombie" servers, running but not delivering valuable results?**

- **Will future software releases even be supported on existing hardware?**

- **Can older servers that do not deliver as expected be repurposed for less demanding uses?**
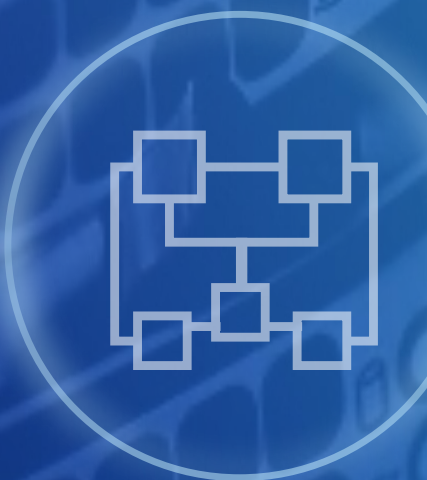
CONSIDER:

Look at current application performance.

Planning, Planning, Planning
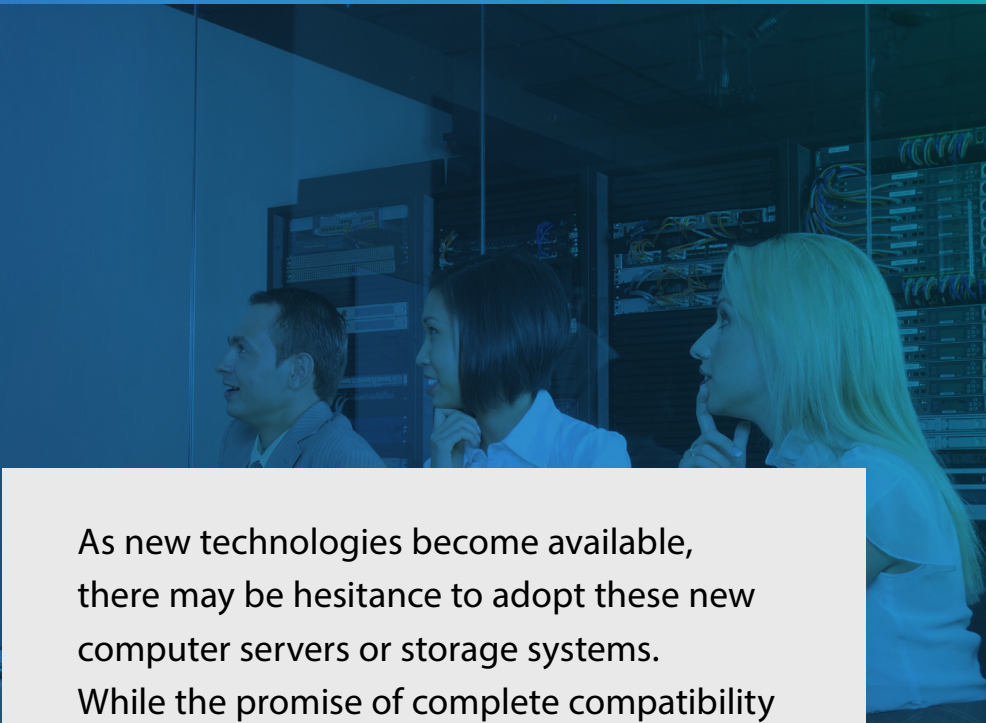
As new technologies become available, there may be hesitance to adopt these new computer servers or storage systems. While the promise of complete compatibility between generations of CPUs and systems is desired for all applications, there may be instances where an entire application ecosystem does not work seamlessly. A process for understanding the ins and outs of a hardware refresh is critical for ultimate success and should be understood. Additionally, new capabilities in new generations of servers, often based on new CPU or GPU technology, can improve performance beyond the increase in the number of cores or clock speed.

CONSIDER:

What new technologies do I need?

Planning, Planning, Planning

New APIs that give access to new features of CPUs or GPUs can also increase performance or security, and these innovations can only be used with the newer systems. Thus, a plan for understanding the new features and modifying applications to take advantage of the new capabilities should be part of any data center refresh.

CONSIDER:

*What software issues
do I need to be aware of?*

Early Testing
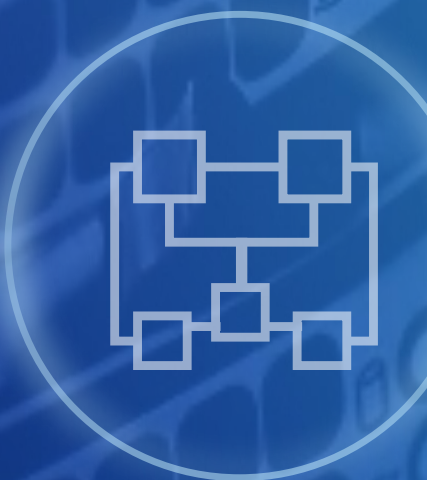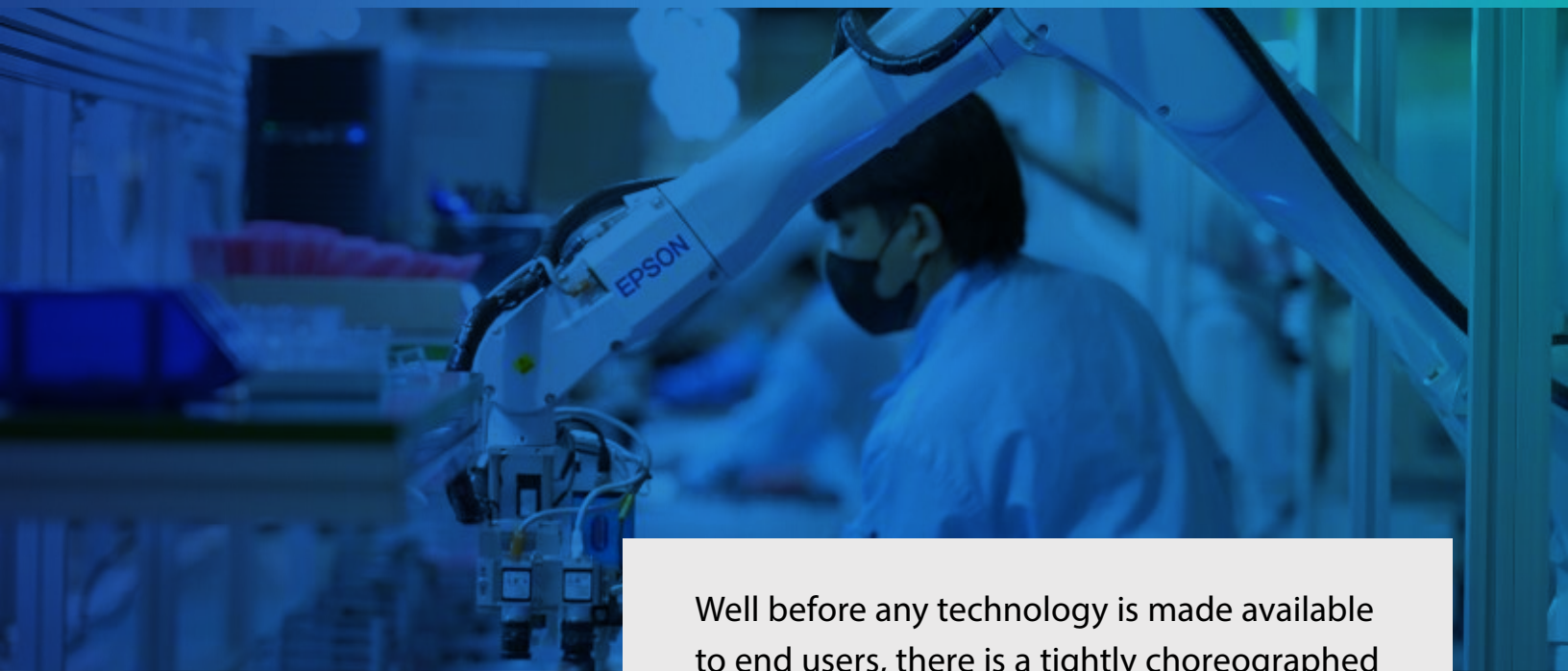
Well before any technology is made available to end users, there is a tightly choreographed process between different suppliers that ultimately results in high quality products. In the IT industry, this starts with the CPU or GPU suppliers delivering the next generation of their products and associated bootup and BIOS settings and code to the server suppliers. This process can be quite long, as modern CPUs and GPUs contain billions of transistors. In addition, significant time is spent running test cases at very low levels and in entire ecosystems.

CONSIDER:

Can I test and validate in advance of a POC or purchase?

Early Testing
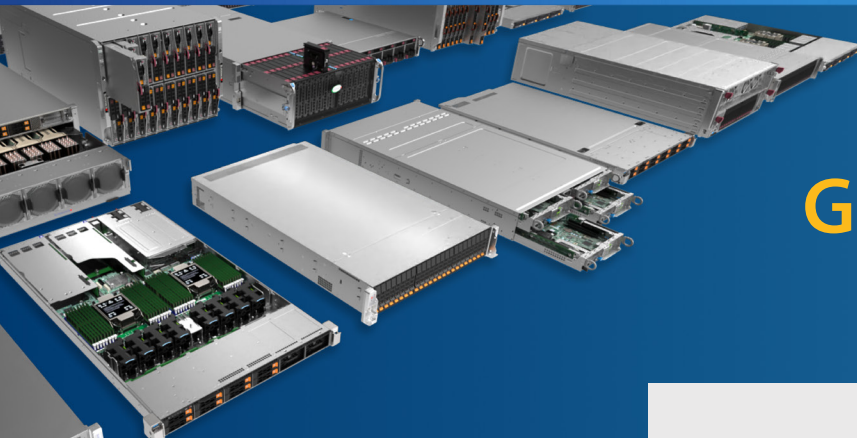
Get Early Access

Before any new server or product line is made generally available (GA), many companies run a program where qualified users can get access to the servers and run actual code. These programs allow developers to test their applications, experiment with different settings and measure results, both for accuracy and performance. Supermicro offers early access to product lines before these servers become available. While the main purpose of making servers available before a product's official launch, Supermicro is continuing to make its JumpStart program available after the recent X13 launch, to developers with a range of servers.

CONSIDER:

Can I use early access to test and benchmark?

Early Testing Systems

With the Supermicro X13 JumpStart program, applications can be installed and tested for compatibility, exploration of new features, and benchmarked. The new systems are designed for workloads, determining which workload should be carefully considered to maximize performance and price and meet SLAs.

CONSIDER:

Which Supermicro product lines am I interested in?

Future Supermicro X13 Lineup

The Supermicro X13 product families include the 4th Gen Intel® Xeon® Scalable processors, with a range of technological innovations. The Supermicro X13 lineup includes:

- **Supermicro X13 GPU Servers**
- **Supermicro X13 Hyper and Hyper-E Servers**
- **Supermicro BigTwin® Servers**
- **Supermicro FatTwin® Servers**
- **Supermicro X13 SuperBlade®**
- **Supermicro X13 GrandTwin™**
- **Supermicro X13 Universal GPU**
- **Supermicro X13 Storage Servers**
- **Supermicro X13 CloudDC Servers**
- **Supermicro X13 WIO Servers**
- **Supermicro X13 Edge Servers**
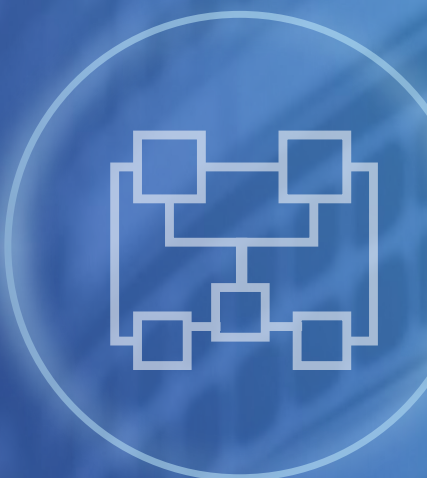- **Supermicro X13 Multi-Processor Servers**

CONSIDER:

Which servers are right for my workloads?
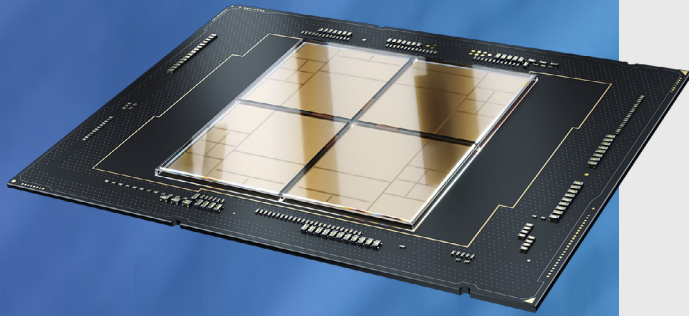
- Intel® Advanced Matrix Extensions (Intel® AMX) extending built-in AI acceleration capabilities of earlier Intel® Xeon® Scalable processors.

- Optimized internal matrix-multiply micro benchmarks run over 7x faster using the new Intel® AMX instruction set extensions compared to a version of the same micro benchmark using VNNI instructions (1).

- 4th Gen Intel® Xeon® Scalable processors will support leading industry standard technologies such as DDR5, PCIe 5.0, and Compute Express Link 1.1.

- 4th Gen Intel® Xeon® Scalable processors integrate the new Intel® Data Streaming Accelerator (Intel® DSA), which is designed to offload the most common data movement tasks that cause the overhead seen in data-center-scale deployments.

- 4th Gen Intel® Xeon Scalable processors will be use the Intel 7 process technology, formerly referred to as "Enhanced SuperFin"

- Support for Intel® Optane™ persistent memory 300 series

- Intel® Xeon® MAX processors, with high bandwidth memory (HBM), integrate HBM to accelerate memory-bandwidth sensitive workloads. HBM will have three modes:
  HBM only, Flat, and Cache.

- Eight channels of DDR5 memory; Four memory controllers supporting eight channels; Integrated memory engine

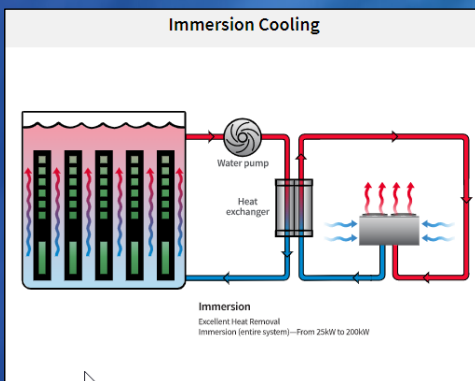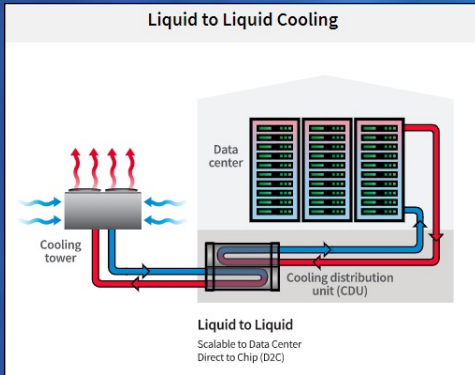- Intel® Ultra Path Interconnect 2.0, with up to 4 links (x24 @16 GT/s)

(1) Claim estimated as of 2/10/2022 based on Sapphire Rapids architecture improvements vs 3rd generation Intel Xeon Scalable at similar core counts on a test scenario using FlexRAN software. Results may vary.

CONSIDER:

What will the 4th Gen Intel® Xeon® Scalable processors give me?

Liquid Cooling

Liquid to Liquid Cooling

Data center

Cooling tower

Cooling distribution unit (CDU)

Liquid to Liquid
Scalable to Data Center
Direct to Chip (D2C)

Immersion Cooling

Water pump

Heat exchanger

Immersion
Excellent Heat Removal
Immersion (entire system)—From 25kW to 200kW

As part of a data center hardware refresh, Liquid Cooling can be accomplished as long as the planning is done in advance. Liquid Cooling of specific servers reduces the data center Power Usage Effectiveness (PUE) in two ways. The first is that the fans in a server that typically draw the cool air over the hot electronics do not have to run as fast since the CPUs or GPUs will not generate as much heat that needs to be removed by flowing air. The second is that the data center HVAC units do not have to be run as much to exchange the hot air for cooler air. Whether Direct To Chip or immersive, liquid cooling reduces the power consumption within the data center but must be planned for in advance to ensure that the proper infrastructure is available.

CONSIDER:

Do I need to consider Liquid Cooling? Will Liquid Cooling reduce my TCO ?

Newer Is Greener
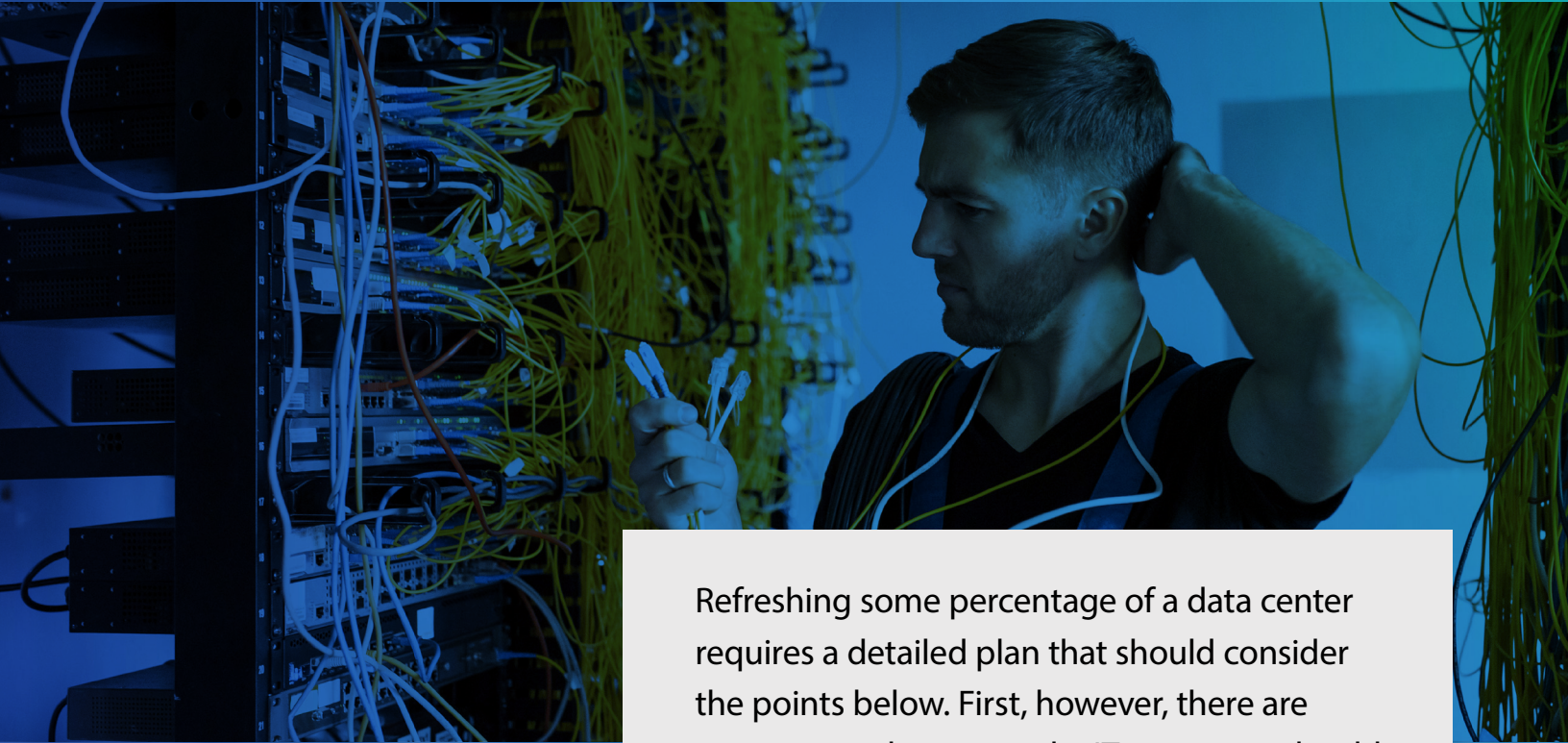
Refreshing some percentage of a data center requires a detailed plan that should consider the points below. First, however, there are some general reasons why IT managers should consider refreshing hardware regularly.

Older servers can't keep up: New workloads may require faster CPUs, more data held in memory, additional cores for processing, and faster data access from disk drives or the network. Nevertheless, the latest generations of servers and CPUs continue to offer overall application performance improvements over previous generations.

CONSIDER:

Which servers need to be updated?

Newer Is Greener

Refreshing some percentage of a data center requires a detailed plan that should consider the points below. First, however, there are some general reasons why IT managers should consider refreshing hardware regularly.

Newer Servers are Greener: The latest CPUs and GPUs perform more work per watt of electricity needed. While the electricity required at the very high end for a CPU generally increases with each generation, the amount of work performed outpaces the increase. This means that less electricity will be used for the workloads compared to previous generations of CPUs.

Data centers benefit by being able to reduce the number of servers dedicated to specific applications while maintaining performance levels. With the added performance of a new generation of servers, some workloads can be delivered on a single server through virtualization or containerization.
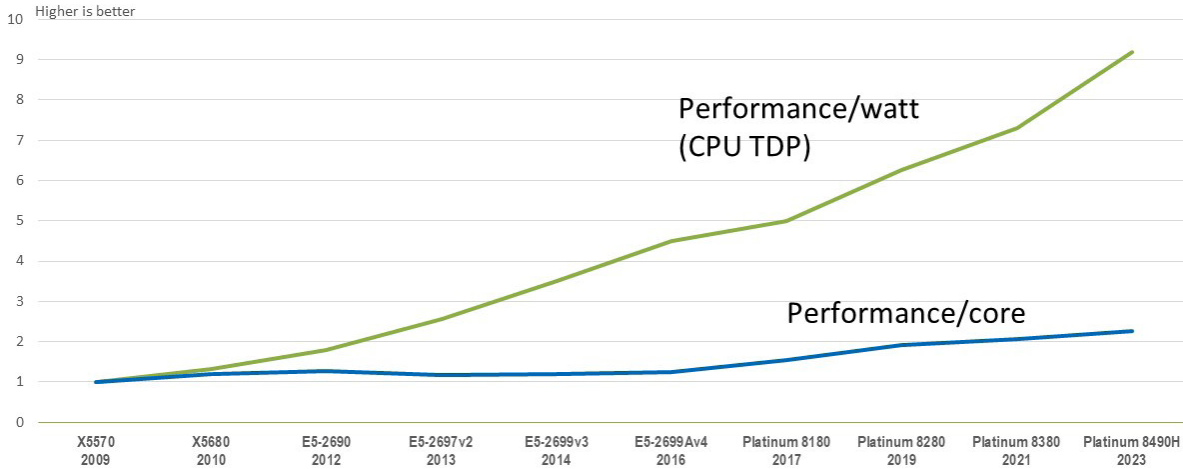
CONSIDER:

Which servers need to be updated?

Increase Your Work Per Watt

Intel® Xeon® Processor
2-Socket Relative Performance/Watt & Performance/Core

Higher is better

Performance/watt
(CPU TDP)

Performance/core

| X5570 2009 | X5680 2010 | E5-2690 2012 | E5-2697v2 2013 | E5-2699v3 2014 | E5-2699Av4 2016 | Platinum 8180 2017 | Platinum 8280 2019 | Platinum 8380 2021 | Platinum 8490H 2023 |

Intel Confidential

Results have been estimated or simulated. Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex. See backup for configurations

intel.
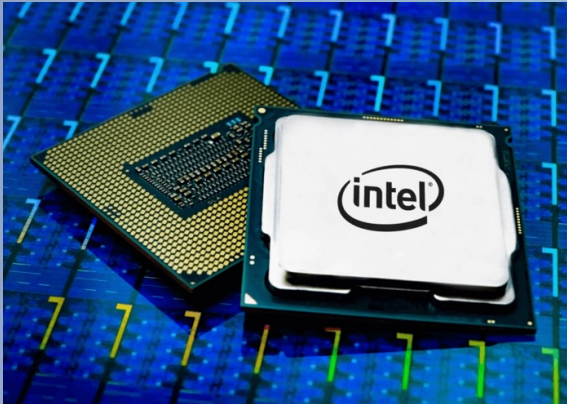
Newer Systems Are More Efficient

Over time, newer systems can perform work per watt of power used. The newer systems are not just faster, but more energy efficient as well.
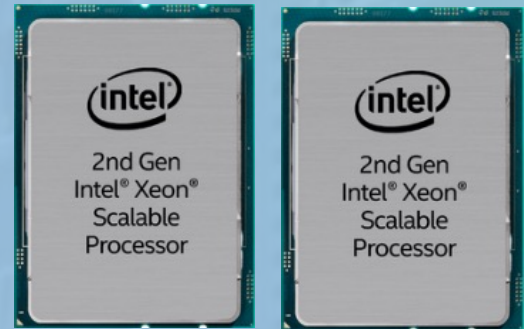
CONSIDER:

Get more work per watt

Newer Is Greener

Newer Generation

>

Previous Generation

Newer systems with the latest CPUs allow for a reduction in the number of CPUs. If workloads allow, a single socket system may be able to replace a previous generation dual-socket system. This not only reduces the number of sockets but could potentially reduce software costs that are tied to the number of sockets.

CONSIDER:

Will a new single socket server allow for the performance that I need?

Environmental Considerations

Data center hardware refresh cycles do not happen all at once. In the 2020 Supermicro Data Centers and the Environment Survey, it was found that 68% of respondents were refreshing their servers within three years. In addition, new systems containing the latest Intel® Xeon® Scalable processors have been designed to perform more work per watt than previous generations. Thus, applications will produce results not just faster than using previous generations of CPUs but with reduced power consumption.
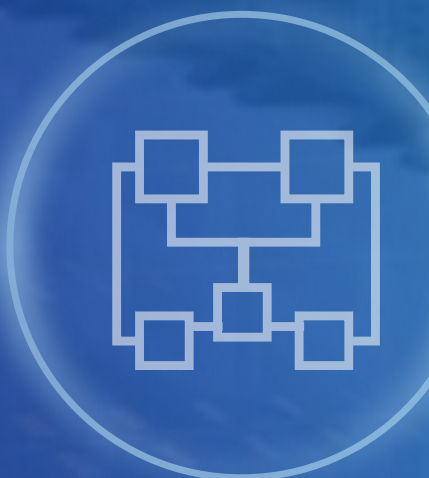
CONSIDER:

How will I reduce my data center environmental footprint?
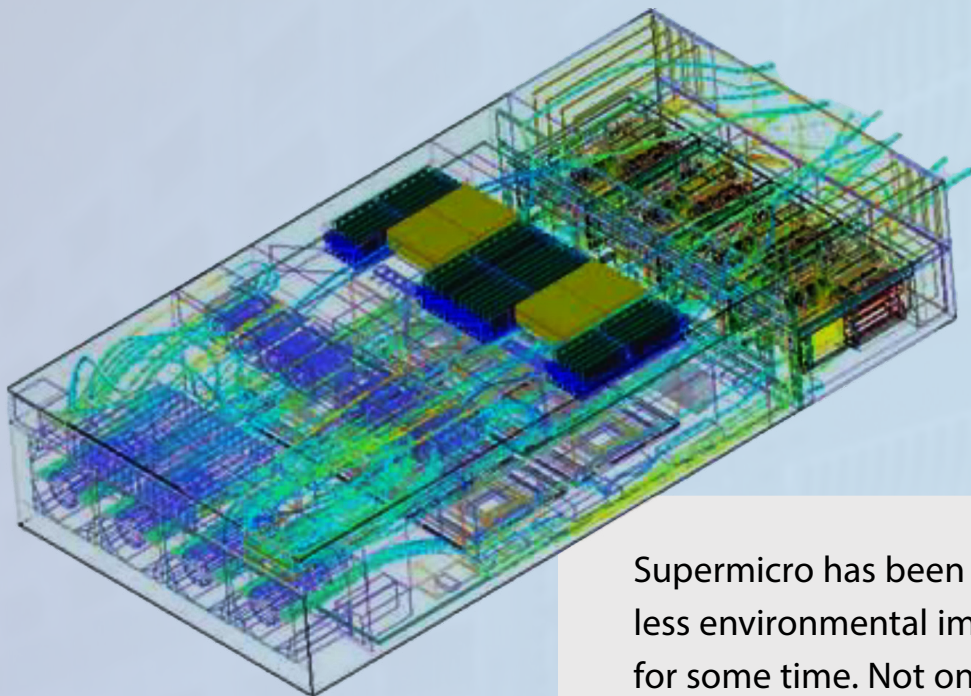
Supermicro has been producing servers with less environmental impact than competitors for some time. Not only do the Supermicro product lines incorporate the latest Intel® Xeon® Scalable processors, but they are also designed through innovative mechanical design to use less electricity. By placing components so that the hottest components receive more airflow, fans may not have to run as fast as other designs, which results in less electricity being used. Also, products such as the Universal GPU are designed to be future-proof, which means that the cooling requirements of next-generation CPUs and GPUs are easily handled.

CONSIDER:

Is energy reduction important to your data center refresh?

Flexibility

Modern data centers need to respond to a wide range of user demands and application profiles. Servers can be optimized for a number of applications, whether the needs are the absolute highest performance, large amounts of memory, high I/O bandwidth to storage devices, or massive networking capacity. Various systems can be optimized for these types of workloads, and these applications' execution will run faster. In addition, specific Supermicro X13 products are designed to give IT administrators the flexibility to use them in different environments within the same product family.
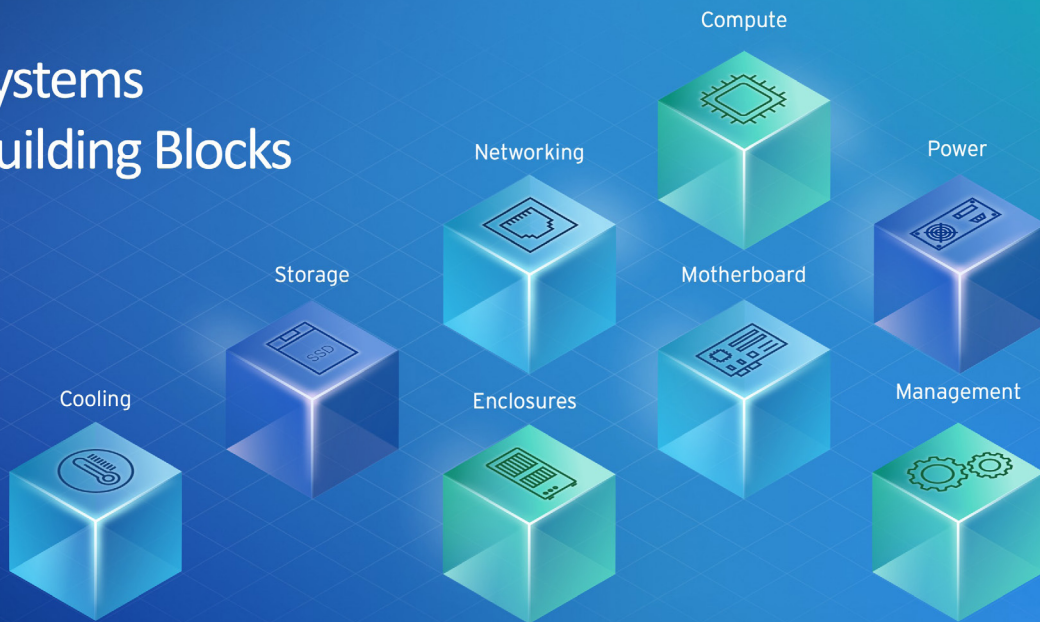
CONSIDER:

Which types of servers do I need?

How To Get Ahead of Your Competitors

Systems Building Blocks

Compute

Networking

Power

Storage

Motherboard

Cooling

Enclosures

Management

Supermicro is able to deliver a range of systems based on the latest technologies soon after introduction. How does Supermicro do it?

By using a concept called "Building Blocks". Many of the systems that together create a state-of-the-art server can be used in different form factors or in application-optimized servers. This design technique allows Supermicro to deliver new servers quickly once new technologies from our suppliers is made available.

CONSIDER:

Getting Systems Fast

How To Get Ahead of Your Competitors

Resource Saving Architectures from Supermicro allow systems to be designed and manufactured that share power and cooling systems. This reduces the electricity used as power supplies are more efficient at more significant loads and larger fans use less electricity when running at a slower speed.

CONSIDER:

Getting Systems Fast

Total Cost to the Environment

$$TCE = f\ (\textit{Server power, power density, inlet temperatures, E-waste activity, PUE})$$

Server environmental considerations go beyond just the electricity used during its useful life. The Total Cost to the Environment (TCE) measures how much a server costs the environment. This is a combination of the effect on the environment during the production process, the cost to operate, and the cost to dispose of the electronic components and mechanical systems. According to a United Nations report in 2020, about 54 metric tons of e-waste were generated. One method to reduce the amount of E-waste and thus reduce the TCE is to design systems that do not have to be entirely scrapped when new components are developed.

CONSIDER:

How can you lower your TCE?

Total Cost to the Environment

For example, with systems designed with a "disaggregated architecture" system, CPUs can be replaced without having to replace the entire chassis. Likewise, newer memory technologies or storage technologies can be replaced without having to replace the CPUs or the memory. It is estimated that a chassis could last up to seven years in some cases and house multiple generations of components. Supermicro's mission is to design the most energy-efficient servers using the latest and most powerful CPUs and GPUs. Using advanced CFD modeling and simulation, Supermicro offers a wide range of servers designed to reduce energy consumption.

Even in small amounts, data center refresh offers the opportunity to reduce the environmental footprint and decrease the TCE. In addition, the latest designs from Supermicro using the Next Generation Intel® Xeon® Scalable processors provide an opportunity to increase workloads while reducing electricity usage.

CONSIDER:

How can you lower your TCE?
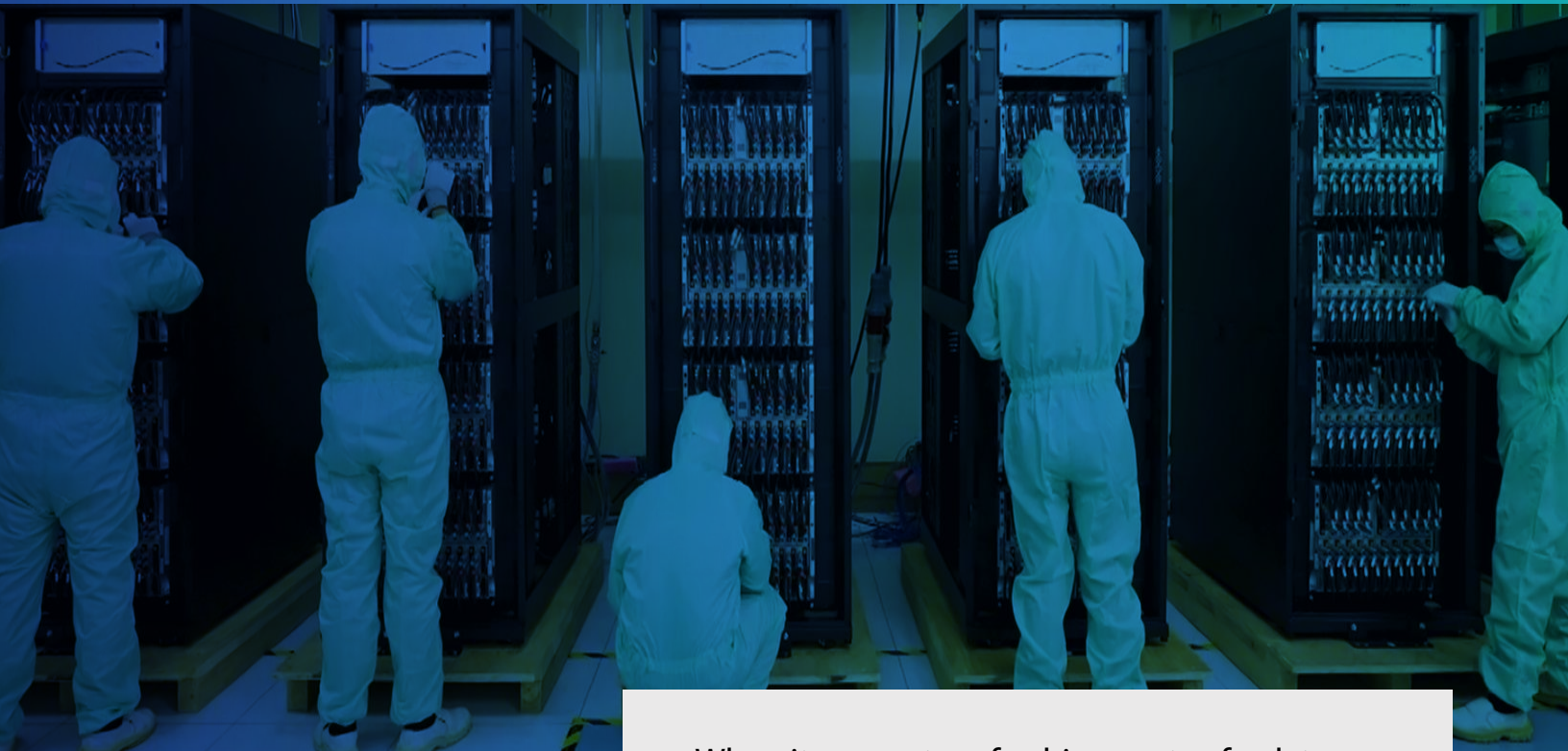
Total IT Solutions

When it comes to refreshing parts of a data center with the most up-to-date servers and storage systems, an entire, ready-to-go, and tested solution is the fastest and easiest method to get productive. Supermicro Total IT Solutions are tested and validated combinations of optimized hardware with software that solve a particular challenge that an end-user organization may have.

**CONSIDER:**

Is a total, delivered, ready to run cluster important to my business?

Total IT Solutions
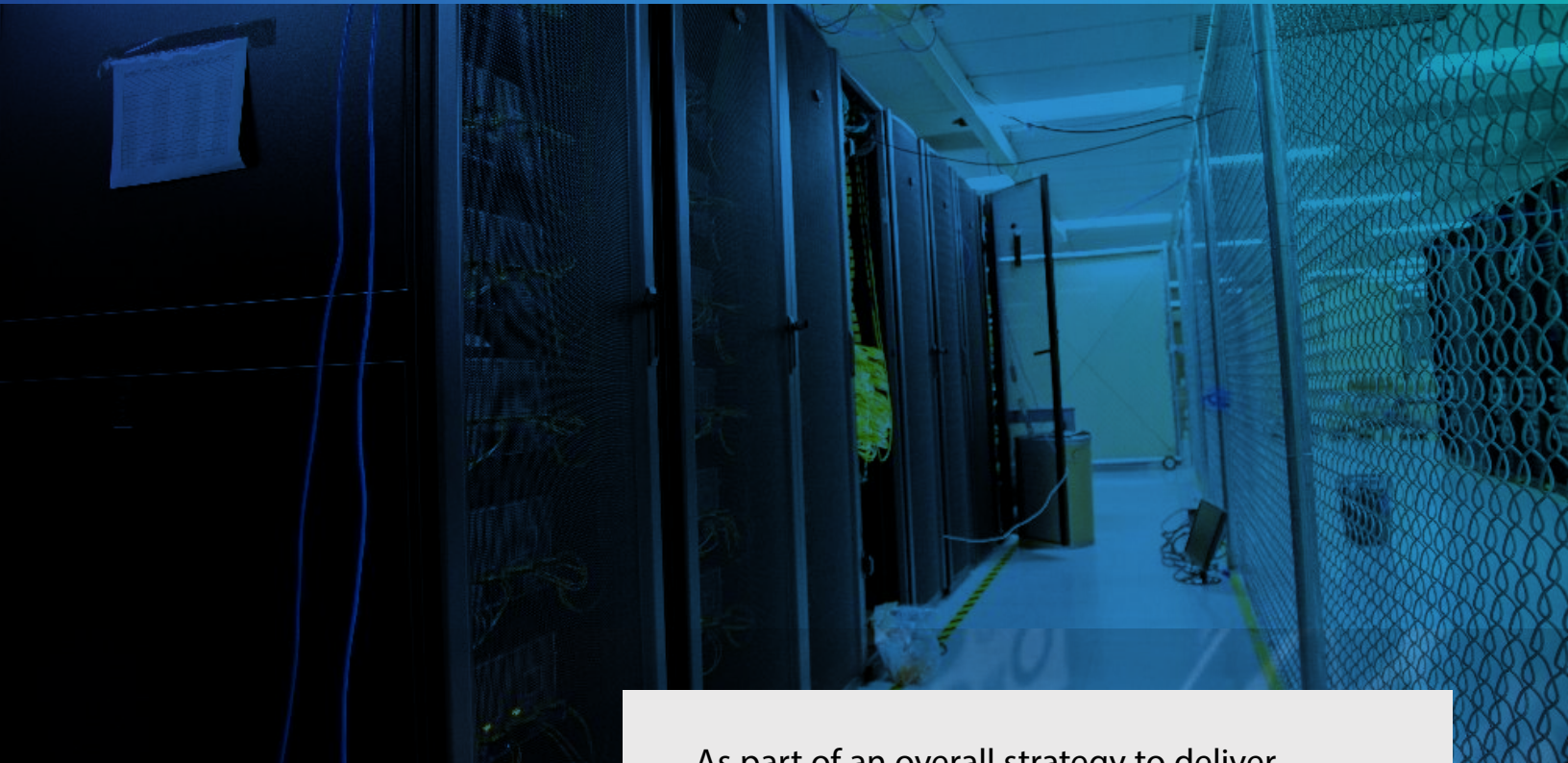
As part of an overall strategy to deliver complete, ready to install racks of systems, Supermicro has developed a range of expertise up to Level 12 (L12). L12 is defined as manufacturing testing at the multi-rack level, where networking is active, tuned, and the entire software stacks are loaded and tested. The whole cluster is considered a single unit and would contain cluster management software and all applications that would be executed.
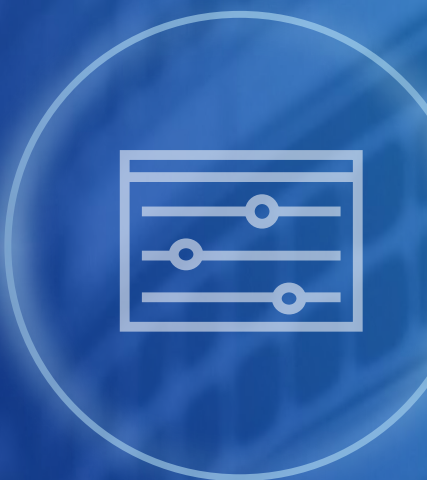
CONSIDER:

Is a total, delivered, ready to run cluster important to my business?

Summary
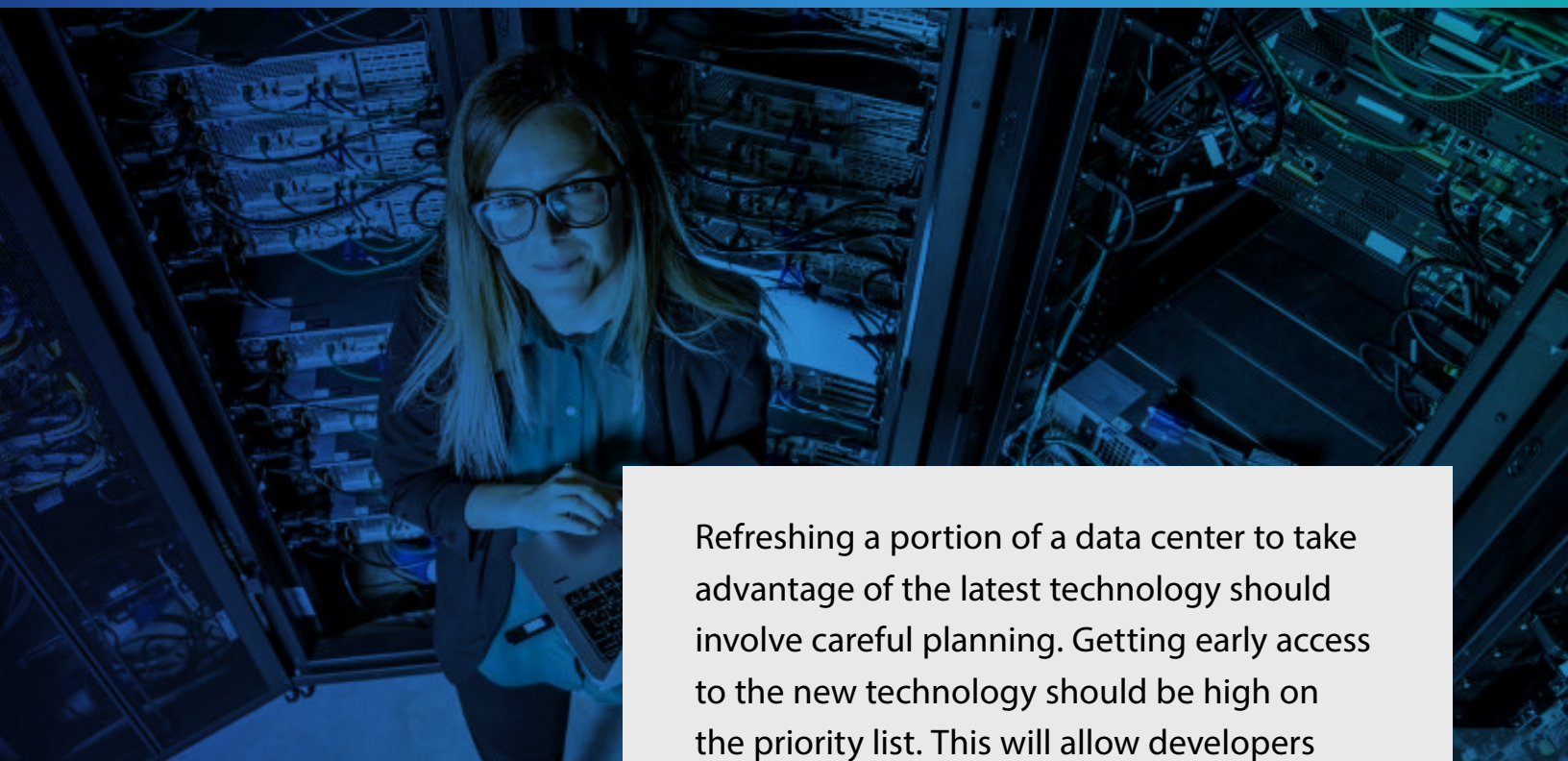
Refreshing a portion of a data center to take advantage of the latest technology should involve careful planning. Getting early access to the new technology should be high on the priority list. This will allow developers and IT administrators to determine if the new offerings will benefit the users of the data center. Supermicro is a complete server and storage solutions provider with advanced manufacturing facilities worldwide. In addition, full-scale cluster-level solutions can be built and tested prior to shipping. Liquid cooling of large numbers of systems simultaneously can also be accomplished to ensure customers get thoroughly tested and validated solutions based on their needs.

CONSIDER:

How smooth can my data center refresh be accomplished?

SUPERMICRO

Learn more at:
www.supermicro.com