

AI Infrastructure Readiness Checklist

Datacenterservices.ai

September 10, 2026

AI Infrastructure Readiness Checklist

What this is

Somebody has approved GPUs. Now the racks, the power, the cooling and the support contracts have to hold them. This checklist covers what changes when you put high-density compute into a data center built for 8 kilowatt racks, and what to check before the pallets arrive.

It is written for the IT director or facilities lead who has to make the existing room work. It is not a vendor comparison. It does not tell you which GPU to buy.

Work through it in order. Each section ends with a yes-or-no question. If the answer is no, that section is your project before the hardware ships.

1. Power per rack

A standard enterprise rack draws 5 to 10 kilowatts. A rack of eight-GPU servers draws 30 to 50 kilowatts at full load. Some liquid-cooled configurations draw more than 100.

Check the following.

The circuit feeding the rack. A 208 volt, 30 amp circuit delivers about 5 kilowatts usable. A 40 kilowatt rack needs multiple 60 amp three-phase circuits, or busway. Count the circuits available at the rack position, not the total room capacity.

The PDU in the rack. Most installed rack PDUs are rated well under 20 kilowatts. Check the plate rating on yours.

The UPS. Add up the new load and compare to the UPS nameplate and the current draw. Many rooms run the UPS at 60 to 70 percent. Forty kilowatts of new load can push it past 100.

The utility feed and the generator. If the room is already close to the building feed, no amount of rack-level work fixes it.

Question: can each GPU rack position deliver the full rated load of the servers going into it, with the UPS and generator covering it, today?

2. Cooling per rack

Air cooling tops out somewhere between 20 and 35 kilowatts per rack depending on containment and airflow. Above that, the servers throttle or shut down.

Check the following.

The cooling capacity of the room in kilowatts, against the new total load. Cooling capacity and power capacity are separate numbers. Rooms are often power-rich and cooling-poor.

Airflow at the rack. Hot-aisle or cold-aisle containment, blanking panels in every empty U, and no gaps under the rack. A GPU server moves a lot of air and pulls hot exhaust back through any gap.

Rear-door heat exchangers. These bolt onto an existing rack and remove 20 to 40 kilowatts using chilled water. They are the fastest way to get an air-cooled room to GPU density without rebuilding it.

Direct liquid cooling. Above 40 to 50 kilowatts per rack, the servers themselves need liquid to the chip. That means a coolant distribution unit, manifolds, and facility water or a chiller loop. It is a plumbing project with a lead time.

Question: can the room remove the heat from every GPU rack at full load, with a stated plan for the racks above 30 kilowatts?

3. Floor, weight and space

A loaded GPU rack can weigh over 1,500 kilograms, or 3,300 pounds. Raised floors and older slab floors have limits.

Check the following.

Floor loading rating at the rack positions, in kilograms per square meter or pounds per square foot. Ask the building.

Rack depth. GPU servers are often deeper than the 1,000 millimeter rack you already have. Check the server spec against the rack and the aisle clearance behind it.

Delivery path. Doors, elevators, ramps, and the floor between the dock and the rack. A 1,500 kilogram rack does not go up a flight of stairs.

Question: will the floor, the rack, and the delivery path take the full weight and depth of the loaded rack?

4. Network

GPU servers move data between themselves at rates that ordinary top-of-rack switching does not handle.

Check the following.

The fabric between GPU nodes. Training workloads usually want 200 or 400 gigabit InfiniBand or Ethernet between nodes. Inference workloads are less demanding. Know which you are running.

Storage throughput. GPUs sit idle if the storage cannot feed them. Check the read throughput of the storage that holds the training data against the aggregate demand of the GPU nodes.

Optics and cabling. 400 gigabit optics and the cabling behind them have lead times and cost. Order them with the servers, not after.

Question: is the fabric between nodes, and between nodes and storage, specified and ordered with the GPU hardware?

5. Support after the sale

This is the section most plans skip.

GPU servers are servers. They fail like servers. The parts are more expensive and the lead times are longer.

Check the following.

The support contract on the GPU servers. What is the on-site response, who holds the spares, and where. A single GPU can cost more than an entire conventional server. Ask what the replacement lead time is.

Spares for the rest of the rack. Power supplies, fans, NICs, drives. The same failure rates as everything else you run, at higher density.

The existing hardware around the GPUs. The storage, the network and the conventional servers that feed the AI workload are often the oldest gear in the room. Their support contracts should not be an afterthought. Third-party maintenance covers Dell, HPE, Cisco, IBM and NetApp gear at 40 to 70 percent below the OEM renewal, and keeps it running 3 to 5 years past the OEM end of service life. Moving the legacy estate to a TPM is a common way to fund the GPU support contract.

Liquid cooling maintenance. Coolant distribution units need filter changes, fluid checks and leak inspection on a schedule. Someone has to own that. Ask who.

Question: is there a named provider, with spares near the site, for the GPU servers, the legacy hardware that feeds them, and the cooling plant?

6. People and process

Check the following.

Who is on call at 2 a.m. when a GPU node fails. Name them.

Who can swap a GPU, a rear-door heat exchanger, or a CDU filter without a vendor visit. If nobody, who is the vendor and what is the SLA.

Change control for the fabric. A misconfigured InfiniBand switch stops a training run. Treat fabric changes like storage changes.

Monitoring. Rack inlet temperature, PDU load, and CDU flow rate should be on a dashboard someone looks at. Thermal problems in a GPU rack develop in minutes.

Question: is there a named person or provider for every failure that stops the GPUs, and a monitor that tells them before the servers do?

The one-page version

Print this part.

Power: circuits, PDU, UPS, feed. Each rated for full GPU load. Yes or no.

Cooling: room capacity, containment, rear-door or liquid plan above 30 kilowatts. Yes or no.

Floor: weight rating, rack depth, delivery path. Yes or no.

Network: node fabric, storage throughput, optics ordered. Yes or no.

Support: named provider, spares near site, for GPUs, legacy gear, and cooling. Yes or no.

People: on-call names, hands-on capability, fabric change control, thermal monitoring. Yes or no.

Six yeses and the room is ready. Anything else is a project with a date.

What to do now

If the support line is the one that is missing, that is the one we can fix this week. We maintain GPU clusters, install and service liquid cooling, and cover the Dell, HPE, Cisco, IBM and NetApp hardware around them under one contract.

Send the equipment list to datacenterservices.ai/get-quote. A priced comparison comes back within one business hour. Or call 888-838-8796.

Datacenterservices.ai is a ThomasTech LLC company. Independent. Not affiliated with any OEM.