Microsoft

Observations on Planning and Operating Very Large Scale AI Infrastructure

Andrew Jones
Principal Product Leader, Future Supercomputing

www.linkedin/com/in/andrewjones
@hpcnotes [.bsky.social] [@mast.hpc.social]

Andrew's observations and comments, not necessarily Microsoft's

$2\ m/s \times 10^8 + 2\ m/s \times 10^8 = 4\ m/s \times 10^8$?

$2\ m/s \times 10^8 + 2\ m/s \times 10^8 = 4\ m/s \times 10^8$?

$2 \times 10^8\ m/s + 2 \times 10^8\ m/s = 4 \times 10^8\ m/s$?

Scale matters (and context)


CNBC
SQUAWK BOX
NADELLA ON $500B STARGATE PROJECT
Subscribe
CNBC
4:16 / 7:32 • OpenAI on Azure
HD

Your DC has 100MW:
How many GPUs can your supercomputer have?

- Assume CSP model says 100MW supply means (e.g.) 75MW assured …
- Assume 1.25 kW per GPU. So it would be 75,000/1.25 = 60,000 GPUs?
- Let's assume 72 GPUs per rack. So that would be 833 racks or 59,976 GPUs.
- Let's assume your deployable stamp is 80 racks. Now it becomes 57,600 GPUs.
- That is 72MW out of the original 100MW … 28MW or 22k GPUs "lost".

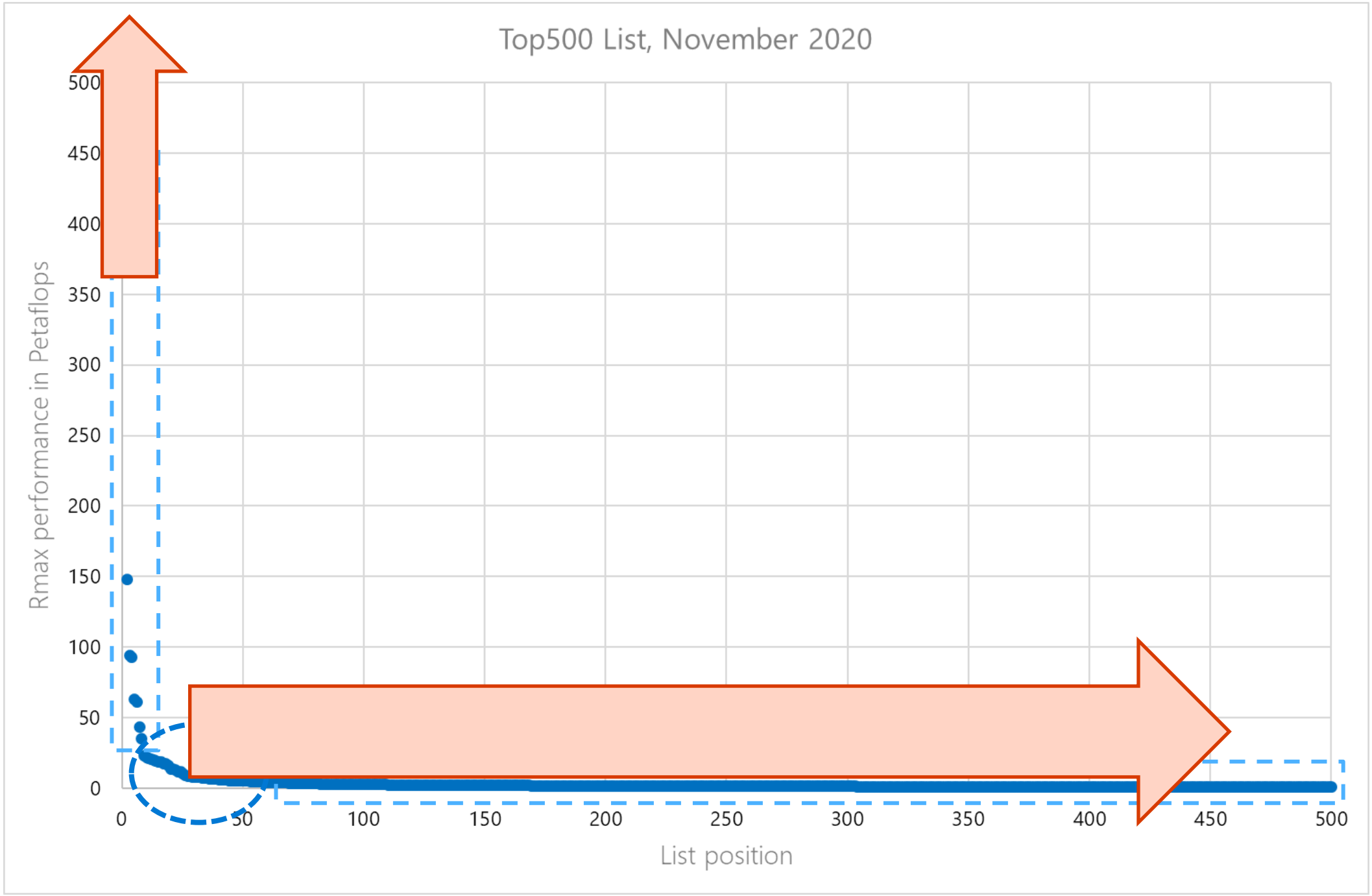
- How much hot spare(s) to carry? That is GPUs or MW or $$ or floor space etc. that are not normally doing useful work. Resilience must be balanced against the global interruptions avoided and thus better overall workload progress.

NB: This is just an example with easy numbers, it is not necessarily indicative of any real GPUs or systems

Your DC has 100MW:
How many GPUs can your supercomputer have?

- What if 100MW supply was used for 100MW at “whatever it is” availability …
- Assume 1.25 kW per GPU. So it would be 80,000 GPUs?
- Let’s assume 72 GPUs per rack. So that would be 1,111 racks or 79,992 GPUs.
- Assume an optimized stamp of just 20 racks. Now it becomes 79,200 GPUs.
- That is 99MW out of the original 100MW … 1MW or 800 GPUs “lost”.
- But lower DC resilience … so which delivers more science / AI training performance? 79,200 GPUs at “good enough” or 57,600 GPUs at high assurance?
- How hard are the constraints? Is it really 100MW? Or 100.8MW = 80,640 GPUs?

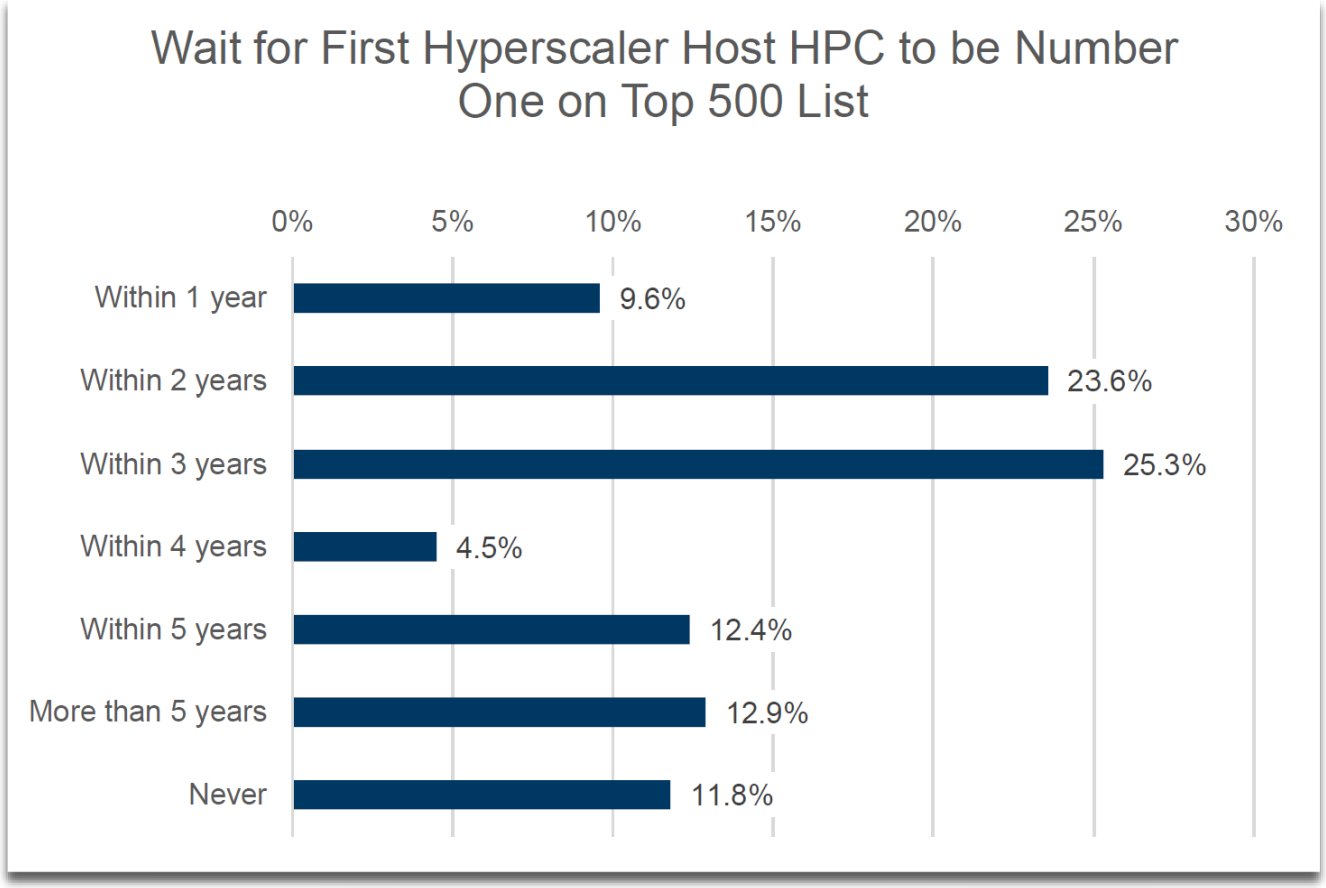
NB: This is just an example with easy numbers, it is not necessarily indicative of any real GPUs or systems


Top500 List, November 2020
Rmax performance in Petaflops
List position
0
50
100
150
200
250
300
350
400
450
500

What is different with many tens of supercomputers?

- Tail end effects
- Technology choices: commonality vs diversity? (reduce or increase risk)?
- Fungibility, repeatable processes, improving processes, …
- What is optimal design point for utilization? How does utilization goals interact with choices of DC architecture, resilience, etc.?
- Varying economic contexts, multi-age fleet, multi-geo, …
- Invest for long term (e.g. software) vs prioritize now (e.g. deployment hacks)?
- How good is good enough? “Within 5%” for $80B is up to $4B!
- First vs best vs confidence vs cost optimal vs revenue optimal?

3) When will a hyperscaler-hosted HPC be the number one system on the TOP500 list?



N = 178, Source: Hyperion Research, 2025

- Three years was most selected option for first hyperscaler HPC to appear in the number one spot on the Top 500 list (25.3%)
- Almost 60% see it happening in the next three years or less
- But one out of 11 don't see it ever happening

Q:
When will a hyperscaler take the #1 on Top500?

- 2x implicit sub-questions:
- (1) when will a hyperscaler have a supercomputer** capable of taking the #1 spot on the Top500?
- (2) will the hyperscaler list that system?
- Answer? A view from the hyperscaler who has a #4 system (was #3) and is maybe the most realistic target of the question ☺ …

Slide withheld ☺

AI has learnt much from HPC. What can wider HPC community learn from AI & cloud worlds?

Earlier science

Is the traditional "all or nothing" model of supercomputer deployment and acceptance (e.g. ready for Top500 run) optimal? Can HPC evolve to multi-step staging of new supercomputers into production for earlier science/business delivery?

Faster science

Sacrificing some headline performance for better reliability often delivers better overall science/business performance.

Quanta matter

Design space is not a continuum. This is not news really, but becomes even more challenging at large scale (or at very small scale). Key feature is that the design space has lots of non-linear pockets of optimization opportunities. Know the relevant performance "curves", etc. Critically, people come in integer quanta too!

Hard constraints

Pure #GPUs per $ or Exaflops/$ are not the best goals or metrics. Better to use maximum science/business value per hard constraint. But know which constraints are hard and which aren't. We often assume what the constraints are (e.g. $) but these assumptions often wrong. Power, space, money, time, risk, your knowledge, your willingness to accept big changes, …

We are interested to collaborate (as peers and/or where the impact is meaningful to both sides).

Microsoft

www.linkedin.com/in/andrewjones

@hpcnotes [.bsky.social] [@mast.hpc.social]