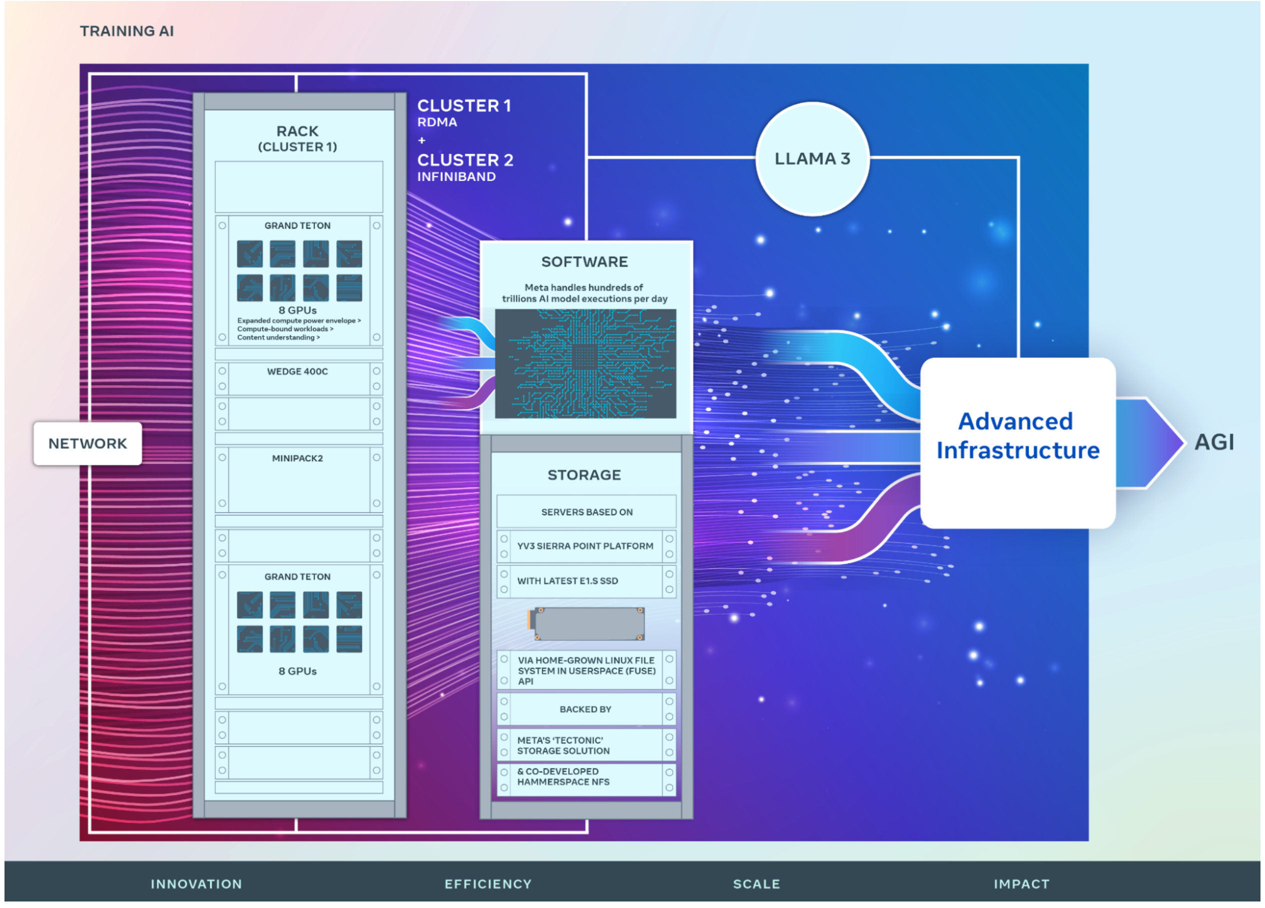
POSTED ON MARCH 12, 2024 TO AI RESEARCH, DATA CENTER ENGINEERING, ML APPLICATIONS

Building Meta’s GenAI Infrastructure



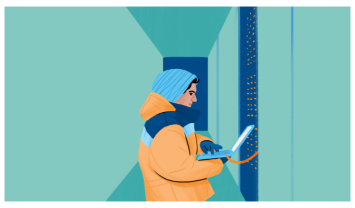
By Kevin Lee, Adi Gangidi, Mathew Oldham

- Marking a major investment in Meta’s AI future, we are announcing two 24k GPU clusters. We are sharing details on the hardware, network, storage, design, performance, and software that help us extract high throughput and reliability for various AI workloads. We use this cluster design for Llama 3 training.
- We are strongly committed to open compute and open source. We built these clusters on top of Grand Teton, OpenRack, and PyTorch and continue to push open innovation across the industry.
- This announcement is one step in our ambitious infrastructure roadmap. By the end of 2024, we’re aiming to continue to grow our infrastructure build-out that will include 350,000 NVIDIA H100 GPUs as part of a portfolio that will feature compute power equivalent to nearly 600,000 H100s.

To lead in developing AI means leading investments in hardware infrastructure. Hardware infrastructure plays an important role in AI’s future. Today, we’re sharing details on two versions of our 24,576-GPU data center scale cluster at Meta. These clusters support our current and next generation AI models, including Llama 3, the successor to Llama 2, our publicly released LLM, as

Related Posts

Jan 11, 2024
How Meta is advancing GenAI



Nov 15, 2023
Watch: Meta’s engineers on building network infrastructure for AI

Oct 18, 2023
How Meta is creating custom silicon for AI

Related Positions

Applied AI Research Scientist, 3D Computer Vision - XR Core-AI
BURLINGAME, US

Research Tech Lead - AI for Chemistry
SEATTLE, US

Research Tech Lead - AI for Chemistry
SAN FRANCISCO, US

Research Engineer - Reality Labs
BURLINGAME, US

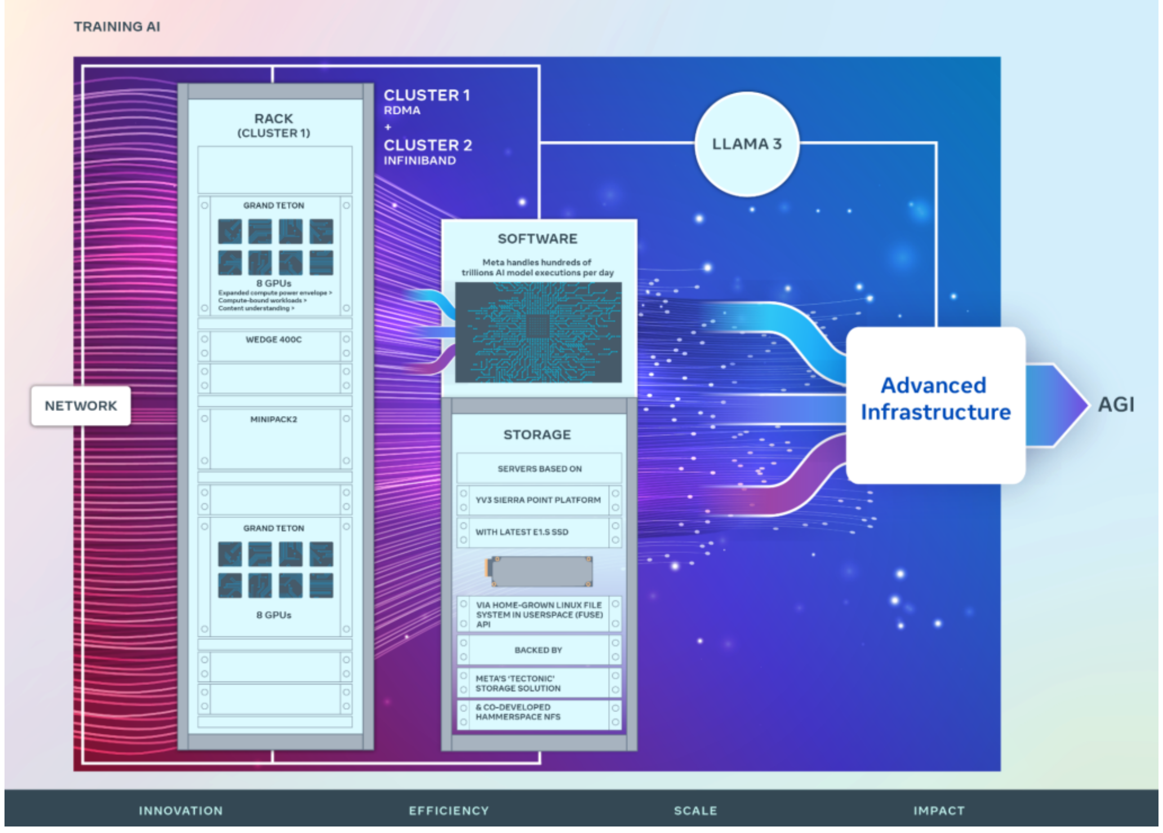
Research Engineer - Reality Labs
NEW YORK, US

See All Jobs

A peek into Meta's large-scale AI clusters

Meta's long-term vision is to build artificial general intelligence (AGI) that is open and built responsibly so that it can be widely available for everyone to benefit from. As we work towards AGI, we have also worked on scaling our clusters to power this ambition. The progress we make towards AGI creates new products, new AI features for our family of apps, and new AI-centric computing devices.

While we've had a long history of building AI infrastructure, we first shared details on our AI Research SuperCluster (RSC), featuring 16,000 NVIDIA A100 GPUs, in 2022. RSC has accelerated our open and responsible AI research by helping us build our first generation of advanced AI models. It played and continues to play an important role in the development of Llama and Llama 2, as well as advanced AI models for applications ranging from computer vision, NLP, and speech recognition, to image generation, and even coding.



Under the hood

Our newer AI clusters build upon the successes and lessons learned from RSC. We focused on building end-to-end AI systems with a major emphasis on researcher and developer experience and productivity. The efficiency of the high-performance network fabrics within these clusters, some of the key storage decisions, combined with the 24,576 NVIDIA Tensor Core H100 GPUs in each, allow both cluster versions to support models larger and more complex than that could be supported in the RSC and pave the way for advancements in GenAI product development and AI research.

Network

advanced and flexible infrastructure. Custom designing much of our own hardware, software, and network fabrics allows us to optimize the end-to-end experience for our AI researchers while ensuring our data centers operate efficiently.

With this in mind, we built one cluster with a remote direct memory access (RDMA) over converged Ethernet (RoCE) network fabric solution based on the Arista 7800 with Wedge400 and Minipack2 OCP rack switches. The other cluster features an NVIDIA Quantum2 InfiniBand fabric. Both of these solutions interconnect 400 Gbps endpoints. With these two, we are able to assess the suitability and scalability of these different types of interconnect for large-scale training, giving us more insights that will help inform how we design and build even larger, scaled-up clusters in the future. Through careful co-design of the network, software, and model architectures, we have successfully used both RoCE and InfiniBand clusters for large, GenAI workloads (including our ongoing training of Llama 3 on our RoCE cluster) without any network bottlenecks.

Compute

Both clusters are built using Grand Teton, our in-house-designed, open GPU hardware platform that we've contributed to the Open Compute Project (OCP). Grand Teton builds on the many generations of AI systems that integrate power, control, compute, and fabric interfaces into a single chassis for better overall performance, signal integrity, and thermal performance. It provides rapid scalability and flexibility in a simplified design, allowing it to be quickly deployed into data center fleets and easily maintained and scaled. Combined with other in-house innovations like our Open Rack power and rack architecture, Grand Teton allows us to build new clusters in a way that is purpose-built for current and future applications at Meta.

We have been openly designing our GPU hardware platforms beginning with our Big Sur platform in 2015.

Storage

Storage plays an important role in AI training, and yet is one of the least talked-about aspects. As the GenAI training jobs become more multimodal over time, consuming large amounts of image, video, and text data, the need for data storage grows rapidly. The need to fit all that data storage into a performant, yet power-efficient footprint doesn't go away though, which makes the problem more interesting.

Our storage deployment addresses the data and checkpointing needs of the AI clusters via a home-grown Linux Filesystem in Userspace (FUSE) API backed by a version of Meta's 'Tectonic' distributed storage solution optimized for Flash media. This storage required for data loading.

We have also partnered with Hammerspace to co-develop and land a parallel network file system (NFS) deployment to meet the developer experience requirements for this AI cluster. Among other benefits, Hammerspace enables engineers to perform interactive debugging for jobs using thousands of GPUs as code changes are immediately accessible to all nodes within the environment. When paired together, the combination of our Tectonic distributed storage solution and Hammerspace enable fast iteration velocity without compromising on scale.

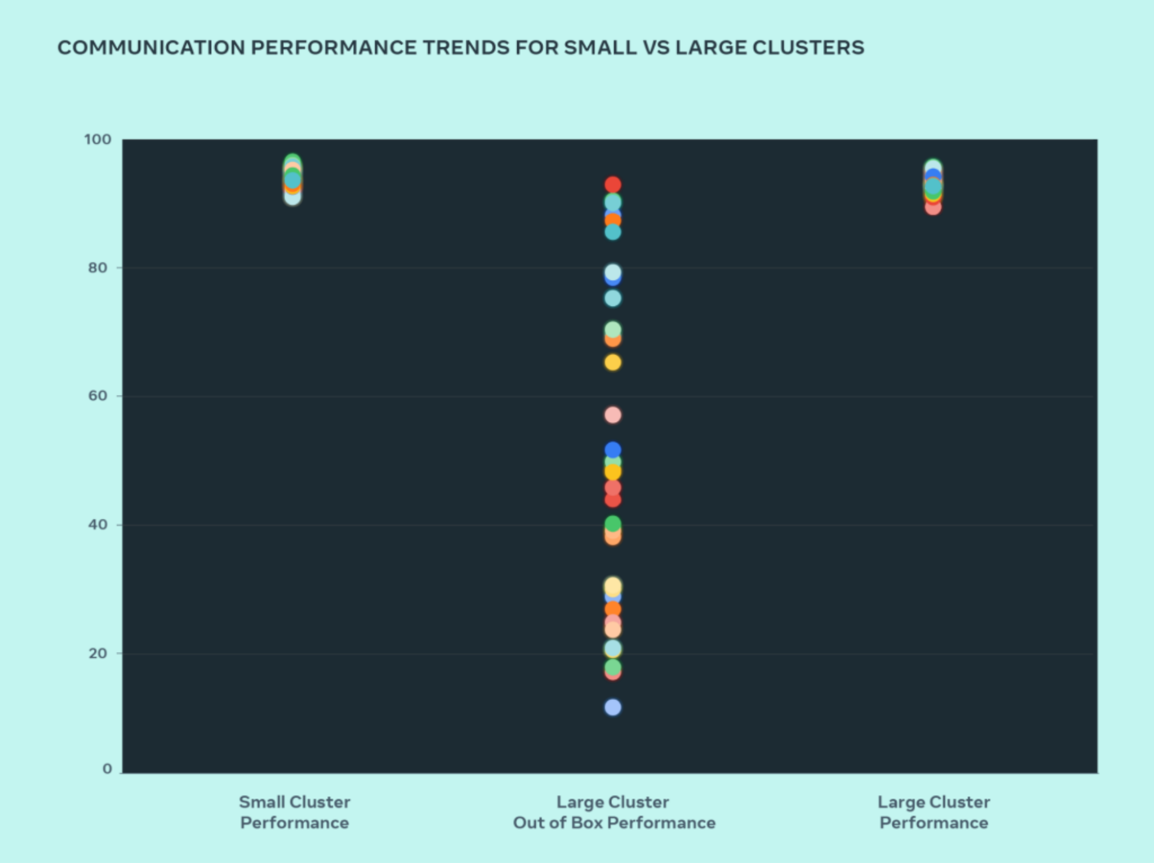
The storage deployments in our GenAI clusters, both Tectonic- and Hammerspace-backed, are based on the YV3 Sierra Point server platform, upgraded with the latest high capacity E1.S SSD we can procure in the market today. Aside from the higher SSD capacity, the servers per rack was customized to achieve the right balance of throughput capacity per server, rack count reduction, and associated power efficiency. Utilizing the OCP servers as Lego-like building blocks, our storage layer is able to flexibly scale to future requirements in this cluster as well as in future, bigger AI clusters, while being fault-tolerant to day-to-day Infrastructure maintenance operations.

Performance

One of the principles we have in building our large-scale AI clusters is to maximize performance and ease of use simultaneously without compromising one for the other. This is an important principle in creating the best-in-class AI models.

As we push the limits of AI systems, the best way we can test our ability to scale-up our designs is to simply build a system, optimize it, and actually test it (while simulators help, they only go so far). In this design journey, we compared the performance seen in our small clusters and with large clusters to see where our bottlenecks are. In the graph below, AllGather collective performance is shown (as normalized bandwidth on a 0-100 scale) when a large number of GPUs are communicating with each other at message sizes where roofline performance is expected.

Our out-of-box performance for large clusters was initially poor and inconsistent, compared to optimized small cluster performance. To address this we made several changes to how our internal job scheduler schedules jobs with network topology awareness – this resulted in latency benefits and minimized the amount of traffic going to upper layers of the network. We also optimized our network routing strategy in combination with NVIDIA Collective Communications Library (NCCL) changes to achieve optimal network utilization. This helped push our large clusters to achieve great and expected performance just as our small clusters.



In the figure we see that small cluster performance (overall communication bandwidth and utilization) reaches 90%+ out of the box, but an unoptimized large cluster performance has very poor utilization, ranging from 10% to 90%. After we optimize the full system (software, network, etc.), we see large cluster performance return to the ideal 90%+ range.

In addition to software changes targeting our internal infrastructure, we worked closely with teams authoring training frameworks and models to adapt to our evolving infrastructure. For example, NVIDIA H100 GPUs open the possibility of leveraging new data types such as 8-bit floating point (FP8) for training. Fully utilizing larger clusters required investments in additional parallelization techniques and new storage solutions provided opportunities to highly optimize checkpointing across thousands of ranks to run in hundreds of milliseconds.

We also recognize debuggability as one of the major challenges in large-scale training. Identifying a problematic GPU that is stalling an entire training job becomes very difficult at a large scale. We're building tools such as desync debug, or a distributed collective flight recorder, to expose the details of distributed training, and help identify issues in a much faster and easier way

Finally, we're continuing to evolve PyTorch, the foundational AI framework powering our AI workloads, to make it ready for tens, or even hundreds, of thousands of GPU training. We have identified multiple bottlenecks for process group initialization, and reduced the startup time from sometimes hours down to minutes.

Commitment to open AI innovation

Meta maintains its commitment to open innovation in AI software and hardware. We believe open-source hardware and software will always be a valuable tool to help the industry solve problems at founding member of OCP, where we make designs like Grand Teton and Open Rack available to the OCP community. We also continue to be the largest and primary contributor to PyTorch, the AI software framework that is powering a large chunk of the industry.

We also continue to be committed to open innovation in the AI research community. We've launched the Open Innovation AI Research Community, a partnership program for academic researchers to deepen our understanding of how to responsibly develop and share AI technologies – with a particular focus on LLMs.

An open approach to AI is not new for Meta. We've also launched the AI Alliance, a group of leading organizations across the AI industry focused on accelerating responsible innovation in AI within an open community. Our AI efforts are built on a philosophy of open science and cross-collaboration. An open ecosystem brings transparency, scrutiny, and trust to AI development and leads to innovations that everyone can benefit from that are built with safety and responsibility top of mind.

The future of Meta's AI infrastructure

These two AI training cluster designs are a part of our larger roadmap for the future of AI. By the end of 2024, we're aiming to continue to grow our infrastructure build-out that will include 350,000 NVIDIA H100s as part of a portfolio that will feature compute power equivalent to nearly 600,000 H100s.

As we look to the future, we recognize that what worked yesterday or today may not be sufficient for tomorrow's needs. That's why we are constantly evaluating and improving every aspect of our infrastructure, from the physical and virtual layers to the software layer and beyond. Our goal is to create systems that are flexible and reliable to support the fast-evolving new models and research.

◂ Prev
Making messaging interoperability with third parties safe for users in Europe

Read More in AI Research

View All ▸