Meta’s Infrastructure Team Needs A Culture Reset

Meta’s Infrastructure Team Needs A Culture Reset Meta Infrastructure has become bloated, with middle managers expending resources on over-engineered technology solutions that lose sight of broader organizational needs.

JUL 21 https://newsletter.semianalysis.com/p/metas-infrastructure-team-needs-a

In our recent newsletter piece about Meta Superintelligence, we expressed reasons to be optimistic on Meta AI. MSL now has many of the right ingredients to catch up with Anthropic and OpenAI to return to the frontier. However, we also briefly alluded to cultural issues plaguing Meta’s infrastructure teams. This article will dive into how these cultural issues have manifested into expensive missteps, whether it be with acquisitions like Rivos or strange choices on hardware architecture.

QQ:1413221435

We believe that Meta Infrastructure needs a cultural reset to better serve the Meta AI organization, especially the world class researchers who are at MSL. This is even more important as Meta embarks on the path of selling its compute to outside customers, not just serving captive internal users.

Meta Infrastructure has become bloated, with middle managers expending resources on over-engineered technology solutions that lose sight of broader organizational needs. The company appears burdened by far too many disparate groups that are over-optimizing for certain metrics as opposed to delivering usable technology for the company as a whole. Middle managers will do everything to justify their proposals to protect their positions within Meta, which has become an extremely political organization.

One big issue is Meta’s six-month performance review cycle, in which the bottom 10% to 15% are cut every review round. The result is an organization of employees that optimize for short-term wins rather than long-term strategies. Some managers push for highly visible projects that can be delivered quickly, a practice known as “window washing” and then promptly pivot or abandon them. Few openly challenge leadership, which leads to bad decisions going uncorrected. The whole system discourages long-term thinking and leads to risk-adverse behavior.

Within Meta Infrastructure, supply chain teams also have little say over engineering teams. The result is technology decisions driven by political motivations rather than thoughtful software/hardware co-design for the broader company.

QQ:1413221435

Frequent pivots are also common. And because Meta has a reputation for throwing money at problems and executing at high speed, these U-turns end up becoming more costly

versus other companies that take a more disciplined or conservative approach. Suppliers also lose faith when given design wins that are later cancelled. This has lead to less supply chain prioritization on new designs. Some suppliers favor focusing on Amazon or Google designs due to Meta's frequent reshuffling.

A lot of Meta's issues come from a lack of financial discipline, with managers creating new projects and headcount to fill out and justify the massive budgets given for initiatives like AI. It's a repeat of what happened to Meta's Reality Labs division, where billions of dollars were spent on engineers and R&D until layoffs, partly brought on by the company's pivot to AI, slashed the team and their related projects starting in 2022 and continuing through this year.

Rivos Acquisition

First, one of Meta's most recent missteps has been the more than $2.5 billion it sunk last year into the acquisition of Rivos, a chip startup.

Few inside Meta's chip division have a full understanding of why the company bought Rivos in the first place, and those who championed the deal internally have since gone quiet. The prevailing theory is that Meta had the money, the custom silicon space was heating up, and it was already licensing Rivos' IP for a future chip, so leadership figured it might as well own Rivos' technology outright than let anyone else have it.

With the Rivos acquisition, Meta also gained the ability to bypass partners like Broadcom to manage the manufacturing and testing of their own custom chips, a practice known as customer-owned tooling. However, paying more than $2.5 billion for that privilege doesn't make much sense either given that a COT team could be built from scratch for maybe $100 million a year or more.

Although Meta wanted only Rivos' accelerator and GPU team, the startup's founders insisted on an all-or-nothing deal. Meta bought the entire company and then heavily cut employees in the parts it didn't want.

QQ:1413221435
QQ:1413221435



Source: SemiAnalysis, Rivos

QQ:1413221435

Some former Meta chip employees say the acquisition was led by Meta's silicon chief, Yee Jiun Song, who pushed for the deal against the wishes of some of those under him but has since lost interest. The result is an organization that has rejected the transplant, leaving Rivos staff stuck to fend for themselves. Existing Meta chip managers treated the acquisition as a pool of free headcount, grabbing Rivos engineers to build out their own empires and pulling the original startup apart in different directions until little of it remained intact.

The technology reasons behind the deal have evaporated just as quickly. Part of Meta's original interest was that Rivos had SIMT core IP which was closer to a Nvidia-style GPU architecture in terms of programmability, whereas Meta's own MTIA is SIMD. Yet after the deal closed, Meta cancelled the chip known as Olympus that was meant to use Rivos' GPU IP, because the design was too aggressive in terms of system and package architecture as well as the software being far from ready. Instead, Meta is sticking with its existing chip architecture through MTIA 600.

Meta still wants to develop a chip that incorporates Rivos technology, and it has created a new chip project known as Phoebe that is now scheduled to tape out in 2028. But not everyone inside Meta is optimistic about the project, and some believe it could eventually be cancelled as well. We have been updating our subscribers of the Accelerator Model about all the different architectures, floorplans, and volumes of various custom silicon.

QQ:1413221435

Meanwhile, incoming Rivos employees have gotten a crash course in Meta's internal culture. Unlike more mature hardware organizations like Apple, Meta's hardware roadmaps can change just as quickly as its software roadmaps, with constant pivots,

meaning that something you are working on now could be pointless in six months. Many within Meta's chip division say there is no clear org structure, no clear decision-making, and no clear ownership of which group does what.

A funny example of the disfunction that is approachable to those not in the weeds, Meta doesn't give most employees assigned desks; people who have been there over a year still scrounge daily for a workspace, where sometimes the desk has a monitor, sometimes the monitor has the wrong cable, and sometimes there is no monitor at all. Employees have formally applied for an assigned desk and been denied, even though desks sit empty everywhere.

Finally, the Rivos acquisition has created dissatisfaction among the pre-existing Meta silicon team. Many Rivos engineers joined with higher compensation and titles without having the requisite responsibilities to justify such perks, and they were placed under existing Meta silicon employees. This resulted in decreased morale within the pre-existing team and also with incoming Rivos staff, who lack the power to make decisions. Since the beginning of this year, a number of Meta's silicon engineers been leaving for startups or other well established companies like Arm and Nvidia.

QQ:1413221435

As we first reported for clients, around 30% of the Rivos engineers who joined Meta as part of the acquisition were let go in recent layoffs. Rivos co-founder Mark Hayter has already left. A number of former Rivos people have since departed Meta to join Gerard Williams's new chip startup, Nuvacore, after their first tranche of RSUs vested in May. We believe that Rivos CEO and co-founder, Puneet Kumar, is also eyeing up an exit in a year or two when his shares in Meta fully vest. There is speculation he could join Rosaic Labs, a new chip startup co-founded by Amarjit Gill, an investor in Rivos and longtime collaborator with Puneet Kumar at SiByte, P.A. Semi, Apple and Agnilux.

Grand Teton

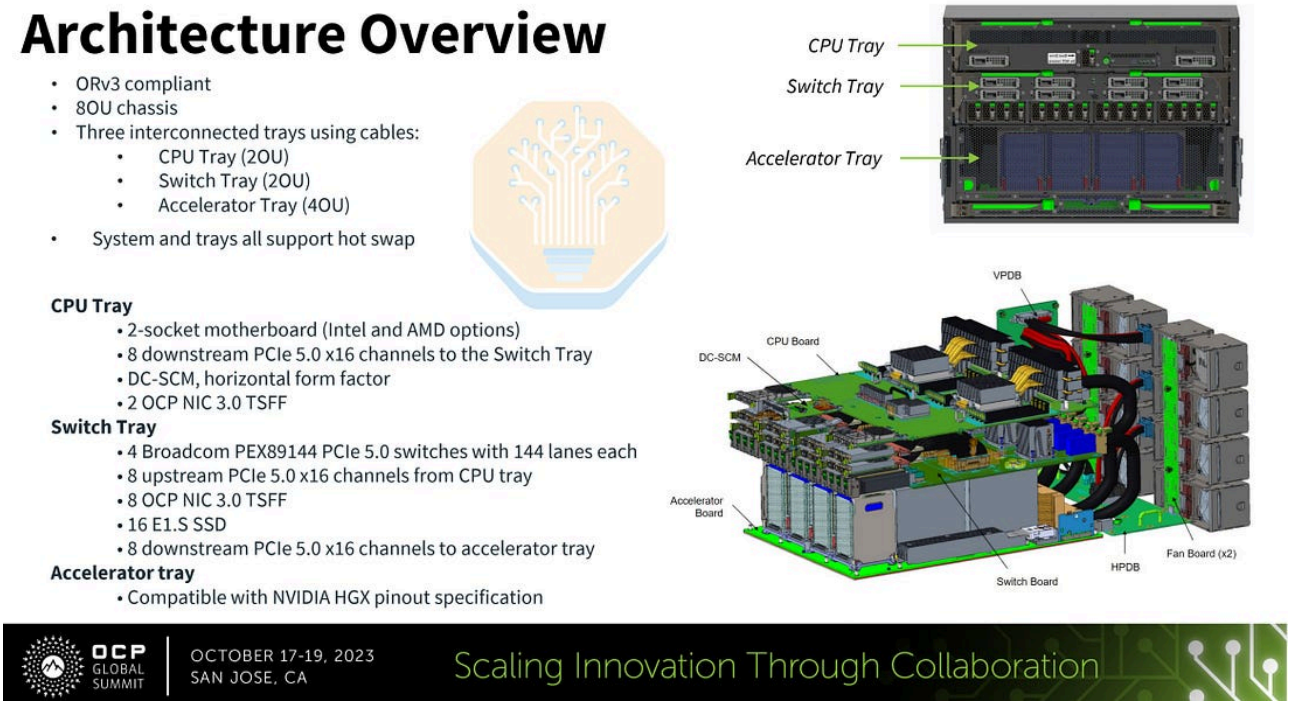
Now let's talk about Meta's AI server designs. Meta's hardware decisions are largely driven by a TCO analysis comparing a matrix of workloads such as ranking-recommendation and GenAI against what vendors like Nvidia, AMD and their own custom silicon offering provide. Meta CEO Mark Zuckerberg has issued an edict that the systems being designed have to work for the entire business, meaning they have to support not only Gen AI workloads but core ranking and recommendation systems as well.

QQ:1413221435

However, Meta has still opted for several "optimizations" for its GPU servers, all of which are strange choices that are often worse than the standard configurations that other hyperscalers purchase. Part of the issue stems from the company's server teams being in completely separate organizations than their networking teams, with different goals and

opinions. Meta engineers are also motivated by the idea of owning their own network operating system and hardware rather than rely on other companies.

To recap, Meta's H100 HGX server was called "Teton Grand" a design that has been contributed to OCP. The major difference from the standard HGX server is the addition of a switch tray in addition to the standard GPU and CPU head-node configuration. This switch tray houses four Broadcom PCIe switches, 16 SSDs and eight NICs. The function is to provide extra PCIe lanes so that the Grand Teton can add more SSDs per server.



Source: Meta

It was during the Hopper generation that Nvidia effectively designed out the need for standalone PCIe switches, by implementing PCIe switching functionality with its ConnectX-7 NIC.

With 80x PCIe lanes available on the Intel Sapphire Rapids host CPU, Meta could have designed a box that connected the 2x 200G front-end NICs with eight lanes each, four lanes for various management and control, and 16 lanes for each of the eight ConnectX-7 NICs. This would leave them 32 lanes to connect to eight E1.S SSDs. So, in the end with Grand Teton Meta was able to squeeze out eight more SSDs per server. This came at the cost of greater server BOM, more power, and more integration complexity.

Why did Meta want all this extra direct attached storage? The Infrastructure team provisioned this with the belief that more storage was needed for checkpointing training runs. However, in production, the storage wasn't utilized nearly as much as anticipated by

the model teams, which is why this design ended up being cancelled. This is one of several examples of lack of hardware and software co-design wasting resources.

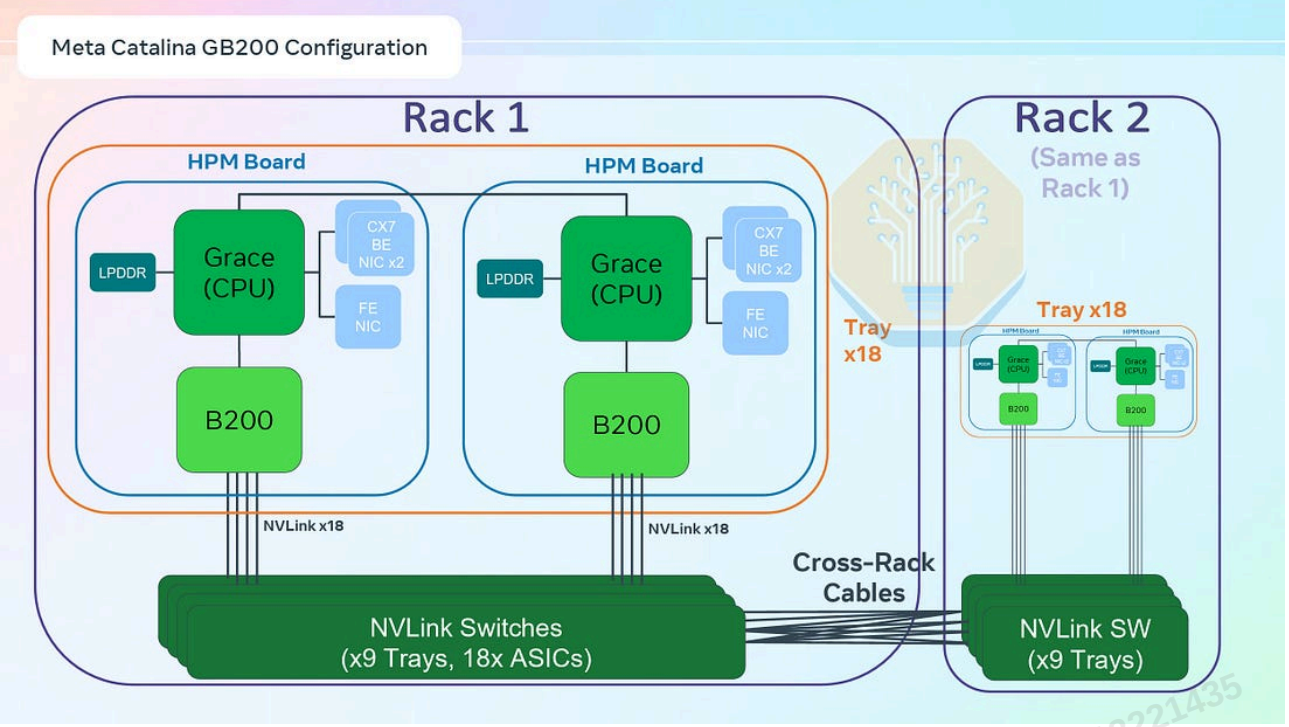
Meta's additional justifications were that the design was easier to service and gave it the flexibility to use non-Nvidia NICs. Its engineers wanted to avoid handing over more business to Nvidia, particularly for networking, given that there already was a strong reliance on Nvidia for GPUs.

But the Broadcom switches couldn't be serviced anyway, and Meta's software stack still depended on Nvidia's version of RoCE, the networking protocol for GPU traffic, so Meta was never realistically going to switch NIC vendors. Ultimately, Meta paid more, had worse TCO, didn't reduce their reliance on Nvidia networking gear, and increased their reliance on Broadcom.

Ariel

QQ:1413221435

The Grand Teton wasn't a one off mistake on server design. This continued into the Blackwell generation with Meta's custom GB200 Catalina rack, also known as "Ariel." Ariel has one Nvidia B200 GPU paired with one Grace CPU, instead of the standard two B200 GPUs with one Grace CPU in the regular GB200 SKU that everyone else purchased. The implementation of this was simple: Ariel used the standard Bianca compute board used for GB200, with each board having one GPU depopulated.

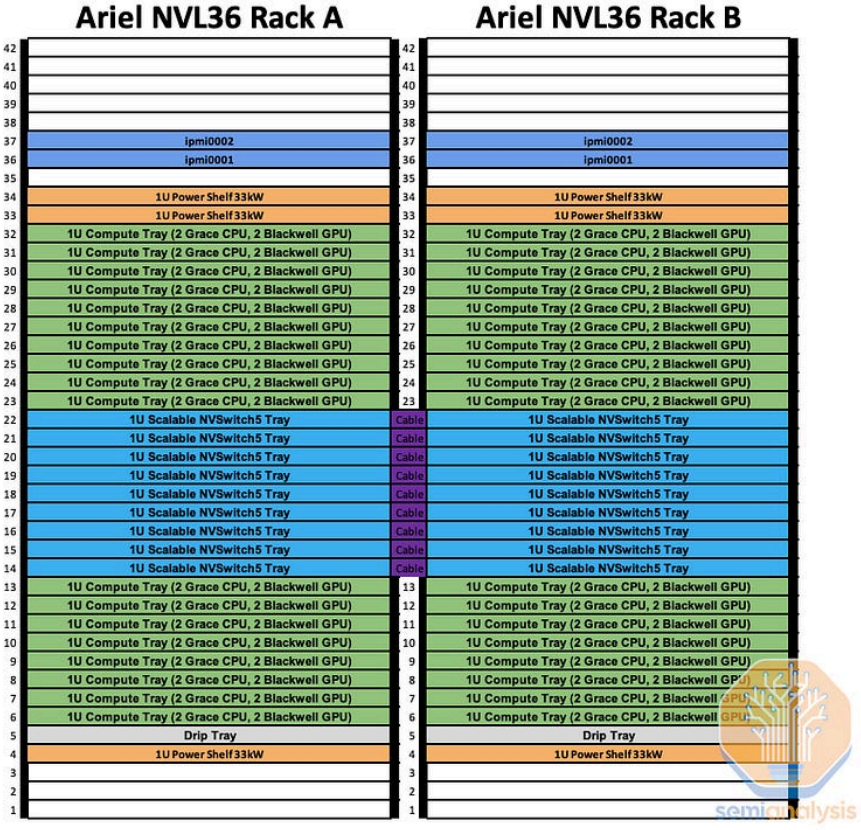


Source: Meta

QQ:1413221435

Meta engineers were concerned about the stability of the single rack GB200 NVL72. NVL72 systems are complex, involving things like a copper backplane that is the central hub where all the chips connect to and miles of dense copper cables and thousands of connections between the chips and the backplane. Meta Infrastructure initially bet that an NVL36x2 configuration would be more stable than running NVL72 entirely through the backplane.

Because Ariel is 36 GPUs in the one rack, to get to the 72-chip world size scale up, it had to be in the 36x2 format. Cross-rack ACCs connected the switch trays between the racks. This meant double the number of switches which while also adding an extra hop of latency.



Source: SemiAnalysis

Meta was the only customer of this Ariel SKU, and the reason for this is Meta's obsession with a higher CPU-GPU ratio for its recommendation systems ("or RecSys') that suggest personalized ads and feeds. The rationale is that RecSys use a lot of embedding tables that store representations of users, ads and content — require a lot of CPU-bound processing and the additional LPDDR memory from each Grace can store the embeddings that have small, scattered reads so bandwidth isn't as important as serving LLMs.

The flipside is that by increasing the CPU ratio and additional NVLink networking content because of the 36x2 configuration the server capex per GPU increases. The result is much

higher $/FLOP and $/HBM capacity and bandwidth than the standard GB200. These are key metrics for LLM/gen AI workloads making Ariel was far worse for LLM training and inference. All in all our calculations show that TCO for the Ariel NVL36x2 server was 14% higher than standard GB200 NVL72. As mentioned this additional expense was for more CPU and DRAM content that isn't being utilized for the LLM teams. This decision cost Meta billions of dollars.

This is not to mention the additional network complexity and reliability challenges that came with the 2-hop scale up with the cross-rack ACCs. These challenges were why a limited number of 36x2 was shipped for some hyperscalers for GB200 only, with the 36x2 configuration being dropped entirely for GB300. Ironically, Meta Infrastructure initially bet that a NVL36x2 configuration would be more stable than NVL72 entirely through the backplane. While the backplane was an initial source of reliability issues, the backplane is now much more mature and won out over the cross-rack cabling. The company underestimated the cross-rack cabline problems with the 36x2 design and overestimated the difficulty of making the backplane work with NVL72. There was no backplane reliability benefit of 36 GPU in 1 rack.

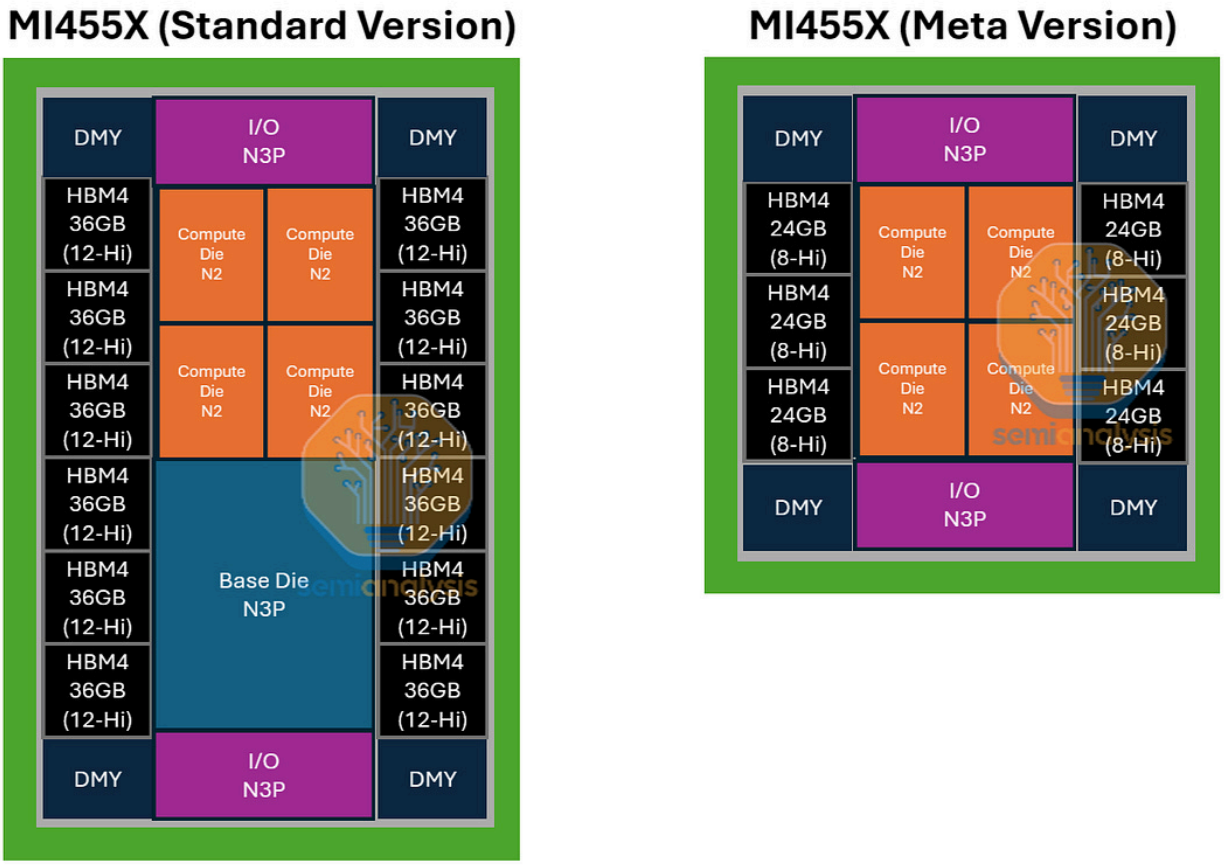
Because we understand that the entirety of Meta's GB200 fleet was in this Ariel SKU, this "optimization" left Meta's LLM teams burdened with an inferior system to the standard SKU that Meta's competitors purchased. This is particularly biting as GB200 was the flagship system for gen AI at the time. Unsurprisingly, there is no more Ariel for Meta's GB300-type servers and Meta is buying the normal configuration.

QQ:1413221435

Upcoming AMD Disasterous Decision

However, we see this now popping up again with AMD's MI450X. In AMD's Q1 2026 earnings call, Lisa Su confirmed that AMD is providing Meta with a custom GPU based on the MI450. We have reported this for many months in the Accelerator Model. This custom Meta MI450 is a cut down version of the full MI450X with half the compute silicon per package and number of HBM stacks. The HBM4 is also downgraded to 8-Hi HBM vs 12-Hi in the standard SKU. Given Meta's Helios rack has the same number of switches, we assume that the amount of I/O is the same.

QQ:1413221435



Source: AMD, SemiAnalysis

QQ:1413221435

By halving the compute and memory silicon, the goal is to achieve a higher CPU compute to GPU compute ratio like Ariel. AMD MI450 will be the most advanced GPU in the market in terms of silicon engineering: 2nm, hybrid bonding, 12 stacks of HBM, with the largest CoWoS reticle size on the market. This results in the highest compute and memory density per chip.

Meta choosing to cut this engineering marvel in half blunts the proposition completely. This chip configuration is intended for recsys workloads and was something that recsys infrastructure teams decided. However, the decision was made before TBD Lab was formed or could have its say. Given the significant compute and HBM deficiency, TBD has no interest in this system, nor is it attractive for external customers.

This decision is going to nuke AMD's volume at Meta because TBD will vastly prefer Rubin if the half MI450 design is chosen. AMD needs to step in, put on their big boy pants, and work directly with teams at TBD to make sure they get the normal MI450 instead of the gimped Meta custom version which is terrible for GenAI. The normal MI450 will actually be competitive with Nvidia's Vera Rubin.

QQ:1413221435

This is a public outcry from us for Meta and AMD to not waste the silicon and have better more efficient and cost effective AI infrastructure. If this change is made we think TBD would actually consider using MI450, driving significantly more traction and business.

So much for Zuckerberg's edict. Next we will go through Meta's elaborate and expensive DSF networking architecture and why Meta quickly pivoted away. We will also discuss how this culture can be fixed.

Network Overengineering

Next is the way Meta developed its networking infrastructure for datacenters. Meta announced its first AI-native networking solution, known as Disaggregated Scheduled Fabric (DSF), in 2024. The goal was to make handling the enormous traffic from model training more efficient and to prevent costly delays in the shuttling of data back and forth between servers.

To better understand the rationale behind the DSF architecture, we must first go back to the problem Meta tried to solve. In large AI training clusters, there are three main issues: elephant flow, low entropy and suboptimal fabric utilization.

QQ:1413221435

> Elephant Flow: AI workloads tend to have long-duration, heavy-traffic flows that can congest the network and reduce the overall performance of the training batch.
>
> Low Entropy: Depending on the training job, the number of IP flows can be limited and congest only a few links while the overall fabric still has plenty of capacity.
>
> Suboptimal Fabric Utilization: Finally, as an overall effect of both elephant flows and low entropy, there is a large skew in the bandwidth utilization of fabric links, which drives how much the fabric needs to be overprovisioned to run smoothly.

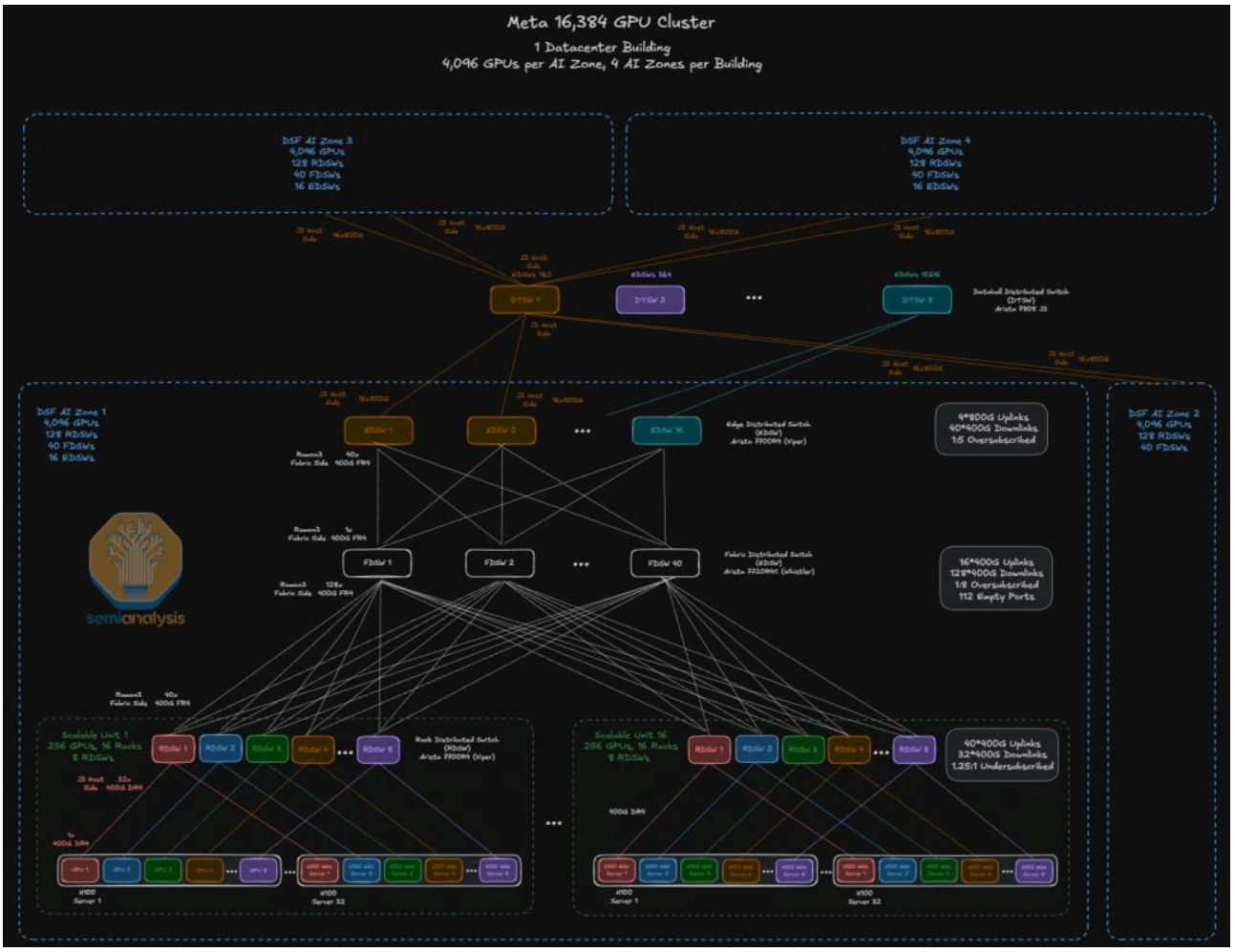
These issues also exist on traditional Ethernet fabrics, but networks using RoCE (RDMA over Converged Ethernet) are turning them from an inefficiency into a failure mode. Indeed, RDMA is highly intolerant to packet loss, and RoCE is built on top of Ethernet, a lossy protocol by nature, and UDP, which is faster than TCP but offers no loss recovery. Software has to be fine tuned through congestion mechanisms to be able to turn the Ethernet fabric lossless.

While RDMA is nice to have on traditional networks, it becomes a necessity in AI networks as AI workloads are latency sensitive and using traditional TCP Ethernet networks would have resulted in a high CPU overhead.

When Meta designed DSF, Network Interface Cards (NICs) were not smart enough to handle congestion control, so Meta decided to tackle the problem at the fabric level. To achieve this, Meta had to go for two different sets of DSF switches, namely the 7700R4C and 7720R4, which are only provided by one vendor, Arista, with effectively a single anchor buyer, Meta, and thus were likely sub-scale when it came to cost structure.

QQ:1413221435

These switches also require the use of specialized switch chips, namely Broadcom's Jericho3-AI, which provides deep buffers to absorb these large and bursty traffic flows, as well as the Ramon3 chip, which acts as the spine switch to forward the traffic sent by the Jericho3-AI chips.



Source: SemiAnalysis AI Networking Model, Meta

At the data-plane level, the two main characteristics of DSF are that it relies on Virtual Output Queuing (VOQ) and cell-spraying. In a naive switch, every input has one queue. If the packet at the head is destined to a congested output, everything behind it stalls, even packets destined to idle outputs. VOQ fixes this part by allocating each ingress its own virtual queue for every egress. In DSF, it means that if one link is congested, packets wait for their dedicated queue while packets destined for other links continue to flow. The other part of DSF, cell-spraying, is a mechanism that breaks a packet into small, fixed-size cells to spray it across multiple links. Packets are then recombined on the other end before being sent to the GPU or XPU. Cell-spraying ensures no link becomes a hotspot.

However, VOQ and cell-spraying aren't enough to maintain a stable scheduled fabric. On top of the virtual queue, a credit scheduler is needed to decide when to drain the queue. Buffering happens at the ingress, and the ingress queue only transmits a packet whenever the egress grants a credit, which is basically the same as saying "I have room for N bytes".

QQ:1413221435

This way, the egress port can never be overwhelmed as it will not grant a credit faster than its own speed. Congestion cannot form at the egress side but only at the ingress due to buffering (which is the reason why Jericho3-AI chips need a large buffer). This whole credit mechanism is the “scheduled” part of DSF.

The control-plane and data-plane complexity were an acceptable tradeoff to Meta at that time as they have their own Network Operating System (NOS), called FBOSS (Facebook Open Switching System). They could have arguably gone for Arista EOS, but FBOSS gave them more flexibility, especially when it comes to the integration of DSF with Meta’s legacy systems, both on the networking side, and on the software side.

To come back on the hardware side, using both Jericho3-AI and Ramon3 ASIC makes the network design inherently complex. For the L1 DSF zone (~ 4,500 XPUs interconnected), two layers are needed, with 7700R4C switches acting as the leaves and 7720R4 switches acting as the spines. The L2 zone (~ 18,000 XPUs interconnected) only needs one supplementary layer of 7720R4 switches to aggregate up to 4 L1 DSF zones. As Jericho3-AI has only 14.4Tbps of network facing bandwidth, it requires two independent fabrics as one plane’s fabric only scales to a certain number of switches before you run out of Ramon3 connectivity. To hit the target GPU count and per-GPU bandwidth in a single AI zone, each NIC/GPU attaches to both planes, and traffic is split between planes at the host/NIC level, which is exactly like plain old ECMP path selection at endpoint, sitting on top of two separated scheduled fabrics, adding complexity to an already complex control-plane.

QQ:1413221435

While DSF was an immediate answer to Meta’s problems around AI clusters, it wasn’t long before the non-scheduled approach caught up. The shift was driven by the NICs themselves getting smarter: as AI clusters scaled through 2023 and 2024, NIC vendors began building RoCE congestion handling directly into their silicon. Once the NIC could absorb the complexity, the fabric no longer had to. Non-Scheduled Fabric (NSF) makes use of switches employing a simpler shallow-buffer switch chip like the Tomahawk 5 ASIC. Virtually every major switch vendor such as Arista, as well as white-box vendors like Celestica, sells switch boxes built around Tomahawk 5, meaning better cost optimization and a diversity of solutions.

As Meta’s GPU clusters grew in size, it became apparent that there were scaling limits to DSF. But by the time Meta engineers realized that NSF was the better implementation, there was no time to develop it given its urgent need of capacity. DSF was deployed with engineers aware that it was a suboptimal and interim solution.

QQ:1413221435

Based on our modeling, DSF is about 11% more expensive than NSF. DSF also forces Meta to rely only on a single vendor, whereas NSF brings vendor diversity. All the above

explains why NSF is now the preferred choice for Meta, with Prometheus and Hyperion deploying a vast majority of NSF instances. Even so, DSF is not going away entirely: which fabric a cluster gets ultimately comes down to which NIC it uses. GB300 deployments run on Nvidia's NICs and use NSF, while systems built around MTIA, Meta's in-house accelerator with its NIC integrated directly on the chip, revert to DSF. The two approaches will co-exist at Meta for the foreseeable future.

None of this means DSF was a blunder. It was the best available answer in 2024, and Meta was smart enough to pivot when a better one emerged.

But Meta's instinct is to throw money at problems and build technically best-in-class solutions at whatever cost, while more conservative hyperscalers took safer routes and were less disrupted when the industry shifted to smarter NICs. Amazon, relentlessly cost-driven, avoided RoCE entirely with its Elastic Fabric Adapter (EFA) running on commodity Tomahawk silicon and providing RDMA semantics. Google treats networking as a software problem, and builds its own transport rather than buying one. Oracle sits in between, building large, functional networks out of simpler hardware without trying to reinvent it. Meta paid for best-in-class, and then the definition of best-in-class changed.

QQ:1413221435

Fixing Culture

The pattern across these different issues is consistent. Rivos, DSF, Grand Teton, Ariel and the cut-down MI450X were defensible on some narrow metric that a group inside Meta Infrastructure was optimizing for, but each was a poor decision for the company as a whole. Meta's model teams inherited worse hardware than what its competitors bought off the shelf, and Meta spent more than they needed to for these solutions.

Hardware decisions need clear ownership and genuine software/hardware co-design with the teams that actually consume the infrastructure, above all MSL and TBD Lab, which should have more of a say over the systems they will train and serve on over the Recsys teams which is often still the tail leading the dog.

Custom SKUs and bespoke fabrics should have to clear a high bar against standard configurations that benefit from industry-wide scale. And headcount and budget should be more disciplined if only to make teams get more creative and conservative about where to put their engineering resources.

This will become more important over time if Meta does intend to sell its compute to outside customers as they likely won't pay for these over-engineered designs. If Meta wants its infrastructure to become a product rather than a cost center, the culture that produced these missteps has to change.

QQ:1413221435

QQ:1413221435

QQ:1413221435