The Fabric of Super Intelligence
AI & Networking Trends (2025-2028)

FMS Santa Clara – Aug 2025

Santhosh Thodupunoori

Sr. Director of Engineering

UpScale AI

Executive Summary – AI Infrastructure Trilemma

- 🚀 **Frontier model scale:** already ≥ 1 T parameters; projections point to 5-10 T (MoE active 50-100 B) by 2028.
- 💸 **OPEX shift:** inference plus RL-based post-training now ~80 % of lifecycle cost for production models.
- ⚡ **Power & IO walls:** racks scaling from 120 kW (2024) → 600 kW (2027) with 1 MW prototypes; requires full liquid cooling and 1.6 T / 200 G-lane SerDes, where reach & signal integrity are the next bottlenecks..

Five Pivotal Shifts (2025-2028)

- **Network-bound era:** MoE & multi-step agents push ≥40 % of run-time into **All-to-All traffic**.
- **Inference OPEX rules:** cumulative serving cost now overtakes training within two years.
- **Post-training surge:** preliminary reports hint Grok-4’s RL phase ≈ 0.5 × pre-train compute (~1e26 FLOPs)
- **Fabric realignment:** closed NVLink / InfiniBand vs open UALink + Ultra-Ethernet (UEC 1.0).
- **Optics roadmap:** 800 G LPO today → NPO pilots (2026) → 1.6 T CPO volume (2027-28).

Strategic Recommendations

- **Co-design** model, RL pipeline & fabric; start board/rack layouts for ≤3″ copper & optical break-out (800 G LPO → 1.6 T CPO).
- **Efficiency first:** deploy 4-bit quant, distill to 20-B MoE, add 2-stage speculative decoding for ≥2-4 × inference speed-up.
- **Bet on openness & light:** back CPO/NPO photonics, open-fabric silicon (UALink-200 & UEC), and asynchronous RL-ops orchestration stacks.

Frontier Models – Mid-2025 Leaders

- 🟣 Grok-4 (≈480 B-param MoE; joint-SOTA MMLU-Pro, SOTA AIME).
- 🟡 Claude-4 (params undisclosed, 200 K ctx); top-tier ARC-C (public score TBD).
- 🔵 OpenAI o3 (128 K ctx, ~15 s TTFT; lowest $/token at $2 | $8).
- 🟢 Gemini 2.5 Pro – flagship reasoning; Deep-Think now on Gemini Ultra.
- 🔒 Llama 3.1 405B + Mixtral-8×22B – open-weight wave drives OSS stack.

Architectural Revolutions

- **Mixture-of-Experts (MoE):** sparse routing unlocks T-param capacity without linear FLOPs, shifting the bottleneck to memory & All-to-All traffic.
- **Structured State-Space Models (SSMs):** Mamba hits **5 ×** Transformer throughput with O(N) scaling; Jamba hybrid delivers **~3 ×** and 256 K ctx in production.

Take-aways for architecture planners

- Expect SSM layers to enter mainstream only when tool-chains catch up (Triton kernels, Flash-Mamba, better KV-cache eviction).
- Near-term (2025-26) the big wins still come from MoE sparsity + quantization + speculative decoding on top of standard Transformers.
- Keep an eye on **hybrid MoE-SSM research**; if the memory/latency claims hold at multi-trillion scale, GPT-6-class models could mix both by 2027.

Network Implications

- **All-to-All traffic is the MoE choke-point:** ≥40 % of runtime on Mixtral-class models; fabric BW/latency is the new throttle.
- **Rail-optimized clusters** (NVLink-Switch pods + leaf-only Ethernet) trim **40-75 % of switch/optic CAPEX** without throughput loss.
- **Towards optical racks:** CPO/NPO deliver **<1 µs intra-rack hops**; 1.6 T (8×200 G) links sample in ’26, volume ’27-28.

Scaling Trends – RL Era

- **Universal toolboxes:** agents now call external APIs by default, yet we’ve benchmarked only a *tiny fraction* of the possible tool space.
- **RL compute surge:** alignment, planning & tool-use loops are compute-hungry; *unconfirmed* estimates put Grok-4’s RL budget at **≈ 50 %** of pre-training FLOPs.
- **Verifiable rewards:** math / code tasks show the power of RLVR, but coverage is still narrow—ample headroom ahead.
- **Debate / consensus methods:** multi-agent critique is early-stage research; most of the solution space remains unexplored.

Inference & RL Efficiency – Key Levers

- **Quantization:** FP4 on Blackwell delivers ~2 × tokens/s vs FP8 at iso-accuracy.
- **Distillation:** 4–20 × model shrink in production (up to 100 × in research) enables on-device LLMs.
- **Speculative decoding:** 2–4 × latency cut; already powering GPT-4o’s “Predicted Outputs”.

Interconnect Debate + Leap to Light

- **Scale-up battle:** proprietary **NVLink-Switch** vs. open **UALink-200 G** (x4 = 800 G), both at 200 Gb/s-per-lane vs **Scale-up Ethernet(SUE)**
- **Scale-out divergence: InfiniBand** (Quantum-III/IV) vs. **Ultra-Ethernet (UEC 1.0)** — an Ethernet-based, RoCE-derived AI fabric. (Debate settled: Ethernet wins!)
- **Optical roadmap:** 800 G **LPO** in volume now → **NPO/OBO** pilots 2025-26 → **CPO** roll-outs 2026-27 for 1.6 T ports.

Hyperscaler Networking Blueprints

- **Google Jupiter:** MEMS optical-circuit switches + SDN let the pod-to-pod topology re-wire on-the-fly.
- **Meta Disaggregated Scheduled Fabric(DSF):** Open Ethernet fabric (Jericho3-AI / Ramon3) under FBOSS & OCP-SAI keeps vendors interchangeable.
- **xAI Colossus:** 400 GbE BlueField-3 SuperNICs per GPU on an NVIDIA Spectrum-X (800 G port) Ethernet fabric.

Strategic Outlook (2025-2028) – Investment Theses

- **Network wall:** MoE & agent systems are now *network-bound* (All-to-All ≈ 30-50 % runtime); progress hinges on high-BW, sub-µs fabrics.
- **Inference gold-rush:** Capital is flooding into HW/SW stacks that cut serving cost (quant → FP4, distill, spec-decode).
- **Photonics boom:** CPO/NPO modules race toward a **$1-2 B** TAM by 2027; entire photonics chain already >$1 T.
- **Open fabrics rise: UEC Ultra-Ethernet + UALink-200G** offer multi-vendor scale-up/out paths and erode proprietary lock-in.

Emerging Bottlenecks (Next 3-5 Years)

- **Power & cooling:** 30-60 MW AI *clusters* drive direct-to-chip and immersion liquid cooling adoption.
- **Data quality ceiling:** marginal gains now hinge on high-density domain data plus filtered synthetic generation.
- **Operational scale:** 100 k-accelerator clusters demand closed-loop AIOps for self-healing networks and firmware.

Action Plan for Technology Leaders

- **Fabric strategy:** weigh NVLink/UALink vs. InfiniBand/UEC on five-year TCO and vendor-flexibility.
- **Infrastructure readiness:** architect racks for optical break-out and direct-to-chip/immersion cooling as loads climb beyond 300 kW per rack.
- **Data-centric ops:** invest in quantisation (FP4), distillation, speculative decoding and high-quality synthetic data to drive down inference & RL cost.

Q&A

- Thank you!