Making the World Safe for Transactions

Open Atomic Ethernet

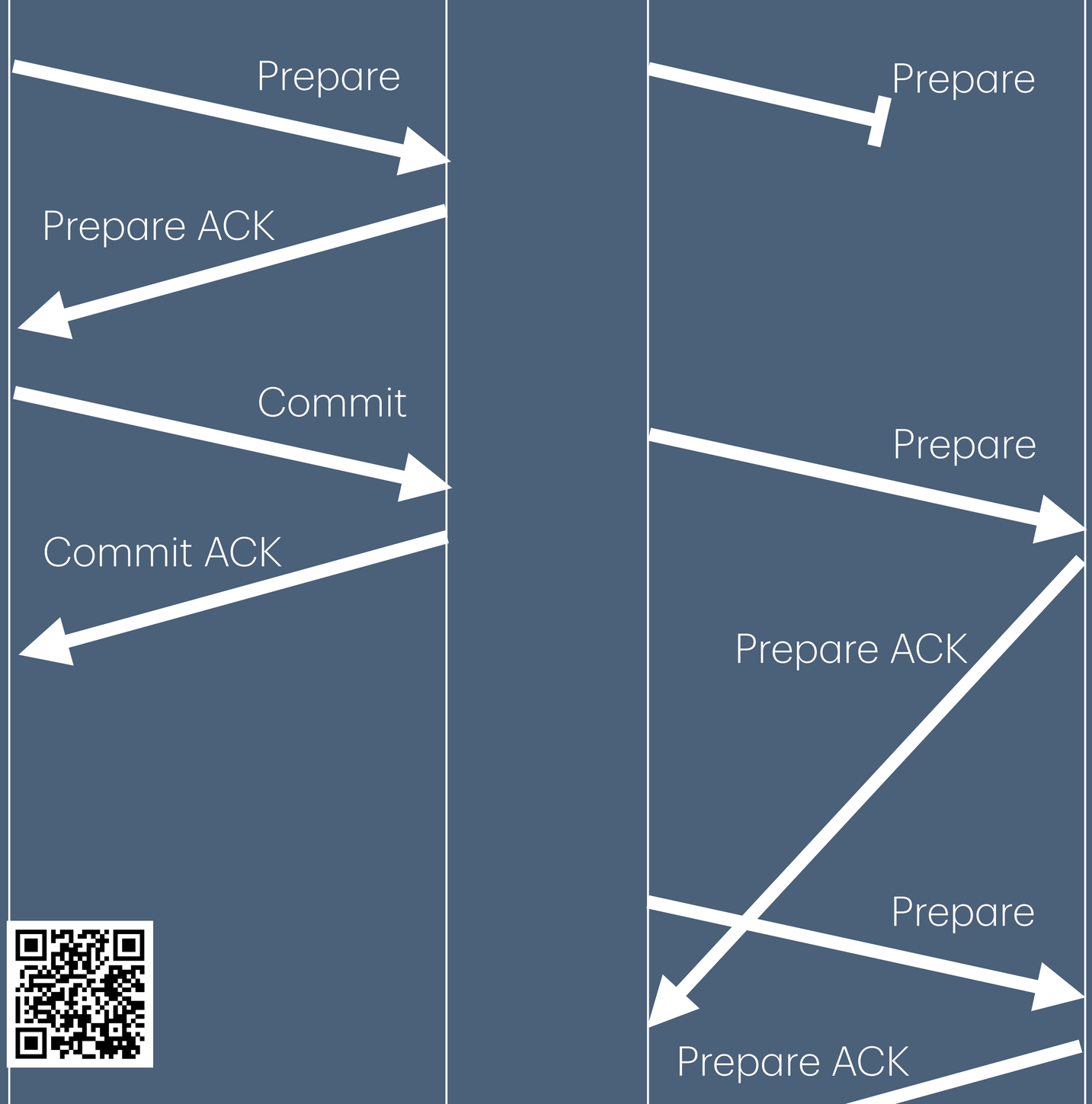
Sahas Munamala 08/07/2025

Why do Transactions Fail

"A transaction is a transformation of state which has the ACID properties: **atomicity**, **consistency**, **isolation**, and **durability**." - Jim Gray

- In local systems, failure is relatively binary: crash or no crash
- In distributed systems, **partial failure**, is a first-class concern

https://xkcd.com/974/



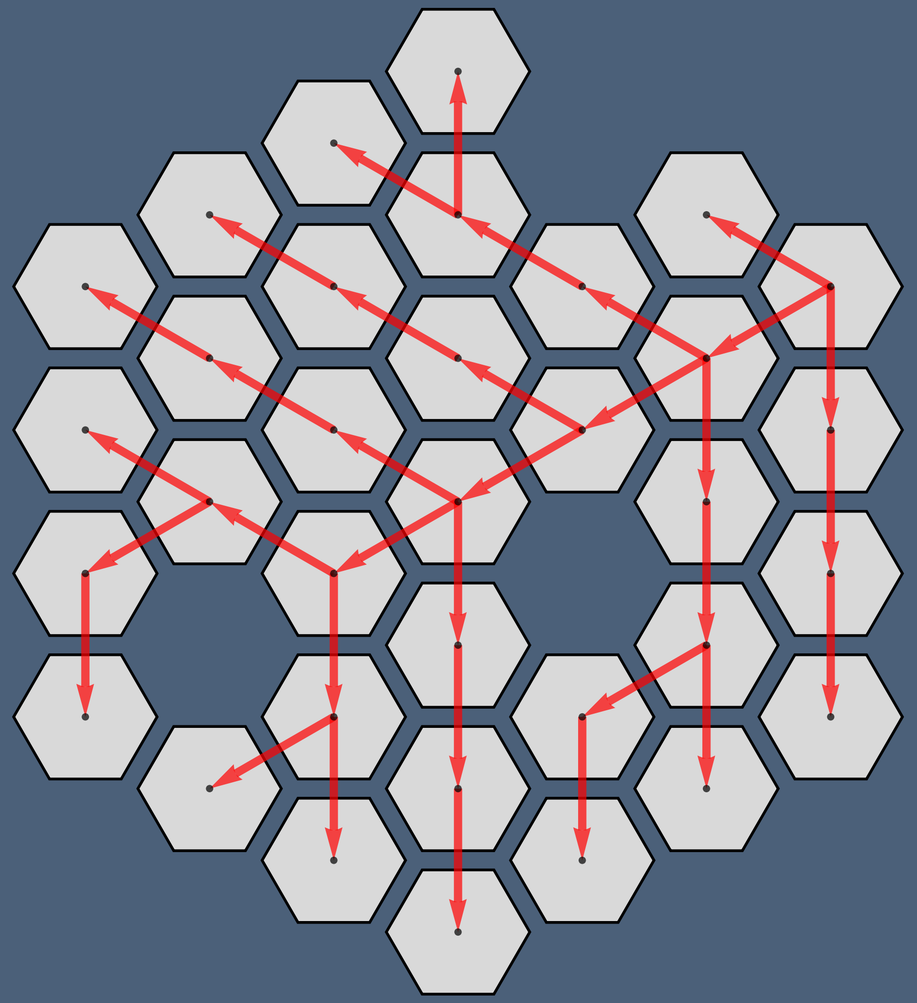
Long tails break consensus and kill throughput.

Is the node dead, or is it slow?

If your only tool is a timeout, then a slow node and a dead node look the same.

A New Contract with the Network

- Atomic Ethernet **guarantees**:
  - In-order delivery
  - Exactly-once semantics
  - Reversible delivery or rollback if reception fails
- Stateless forwarding → deterministic routing
- No dropped packets, no timeouts, no ambiguity



Dimension	NVLink (Gen5 / "Fusion")	UALink 200G1.0	Æthernet (OAE Spec v0.6)
OSI anchoring	PHY + "link-layer-plus" (slot between PCIe & memory bus)	PHY + link-layer; rides its own switch silicon	Inside IEEE802.3 PHY, but rewrites the MAC to be transaction-aware Layer2
Governance / IP	INVIDIA-licensed; Fusion opens ports to 3rd-party silicon but terms remain NDA	Open consortium (AMD, Intel, AWS, Google,...) with royalty-free spec.	AEIA (Atomic Ethernet Industry Association) + OCP permissive license + IEEE (Future). No royalty for spec + Implementation
Peak link BW	18 × 100 GB/s bidirectional per Blackwell GPU → 1.8TB/s in+out	→ ~3.2Tb/s raw	200 G/lane PAM4; x16 lane device today Re-uses commodity Ethernet SerDes (100 G-1.6 T); BW follows optical roadmap
Latency class	~10 ns hop (GPU-NVSwitch)	Sub-100 ns end-to-end target in spec	Aims at single-digit us wire-to-wire deterministic latency with explicit ACK/NACK
Memory model	Cache-coherent load/store (GPU→GPU, GPU→CPU)	Non-coherent memory-semantic load/ store + atomics	Message becomes an ACID transaction; app-level atomicity not cache primacy
Pod scale before routing	NVL72 rack (72 GPUs) today; Fusion gateway bridges multiple pods	Up to 1,024 accelerators per "AI pod" in 1.0 spec. (12 Bit src/dst addresses)	Scale Independent - Planar Cell and Link Spec 3 x 3 = 91-hop physical cell connectivity. 5 × 5 =25 2-hop connectivity
Topology hardware	NVSwitch (IC) backplane or copper cable; Fusion adds external retimer/ bridge	Dedicated UALink switch ASICs; leaf spine fat-tree reference	None: Direct Connect Network. Does not require switches. reliability is in the frame dialect, not the fabric
Failure philosophy	Best-effort delivery inside a fault-contained chassis; relies on upper layers for retry	Same; link-level CRC + retry, no explicit Link-layer ACK/NACK, idempotent ops transaction semantics	timeout-free; wants to remove retry storms
Typical deployment scope	Inside DGX, GB200 NVL72 rack, soon third-party CPU/accelerator	Rack-scale AI training podsfrom multi-vendor OEMs	Data-centre-wide low-latency transactional fabric; also targets SmartNIC & DPU offload

Questions?

Learn More about Atomic Ethernet