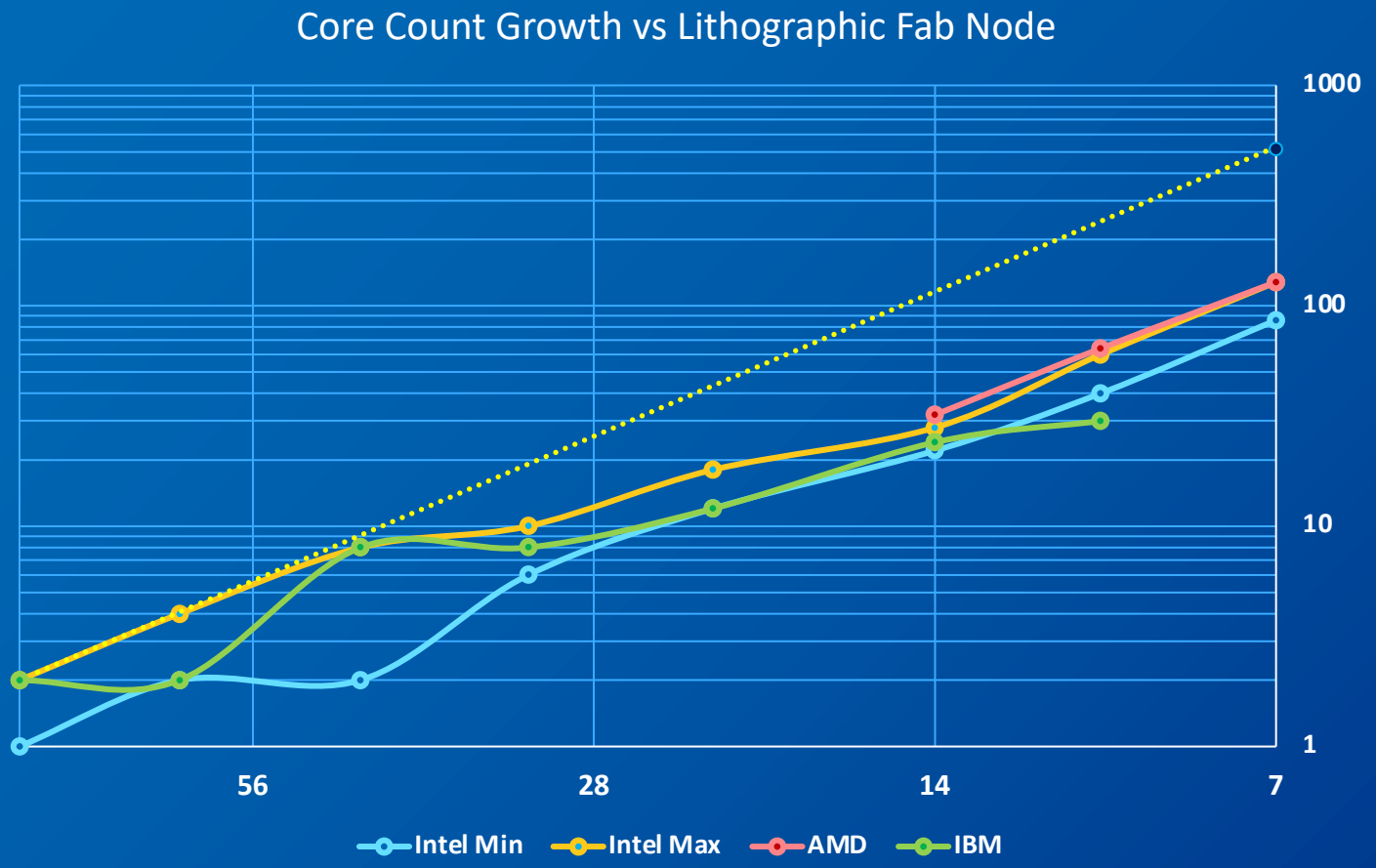
Tomorrow's Shared-Everything Architectures

Josh Fryman, PhD
Office of the CTO, Intel Fellow

05/08/2023

Market Pressure: The Race

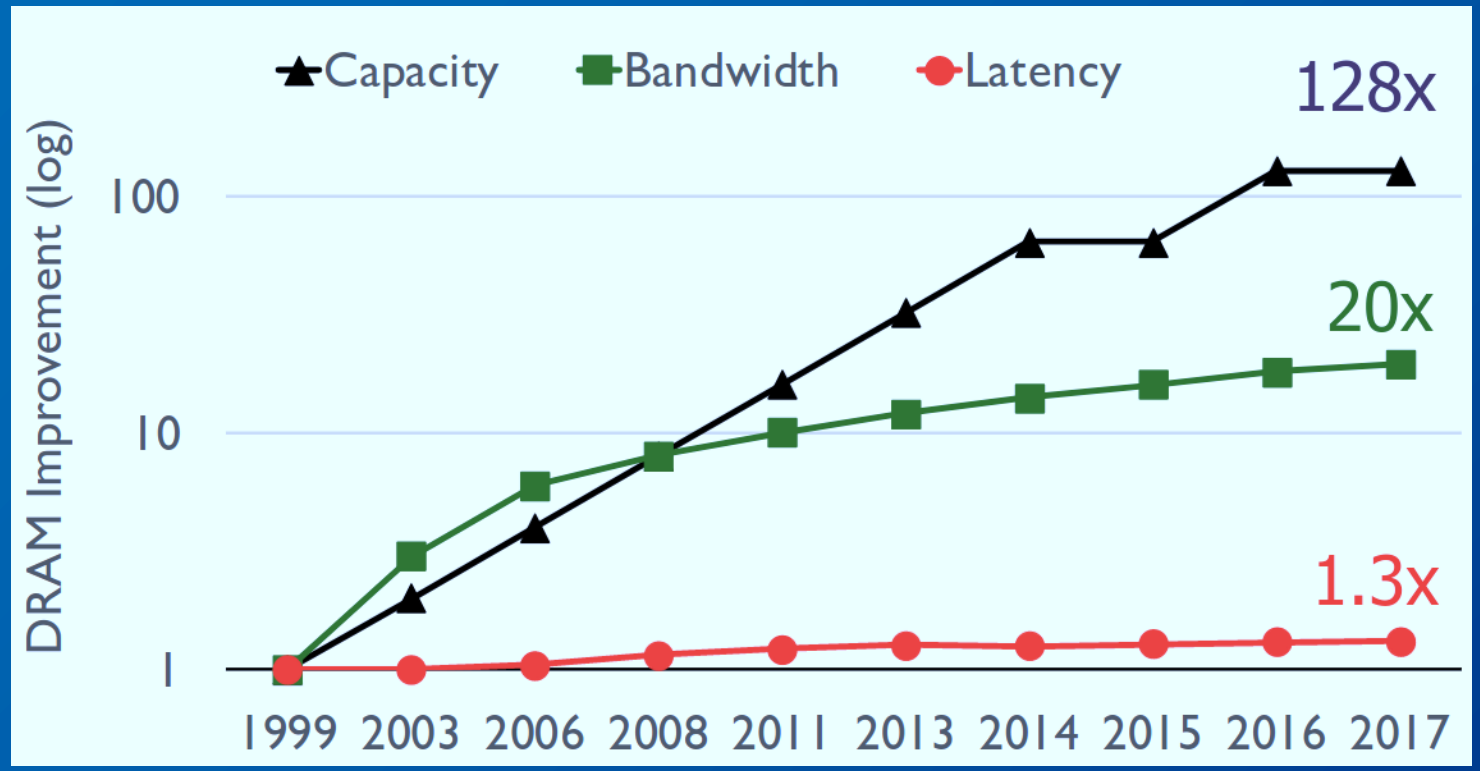


Source: Wikipedia, 2021

Memory capacity/core is relatively constant by market segment . . . 0.5-4 GB/core

Memory bandwidth per core is also relatively constant by segment . . . 2-16 GB/s/core

Market Pressure: The Cliff



Source: "A modern primer on processing in memory," by Onur Mutlu et al, arxiv.org, Dec 2020.

Recent Historical Prices

DDR memory typically in $3.5-4.5/GB range

HBM memory typically in the $8-10/GB range

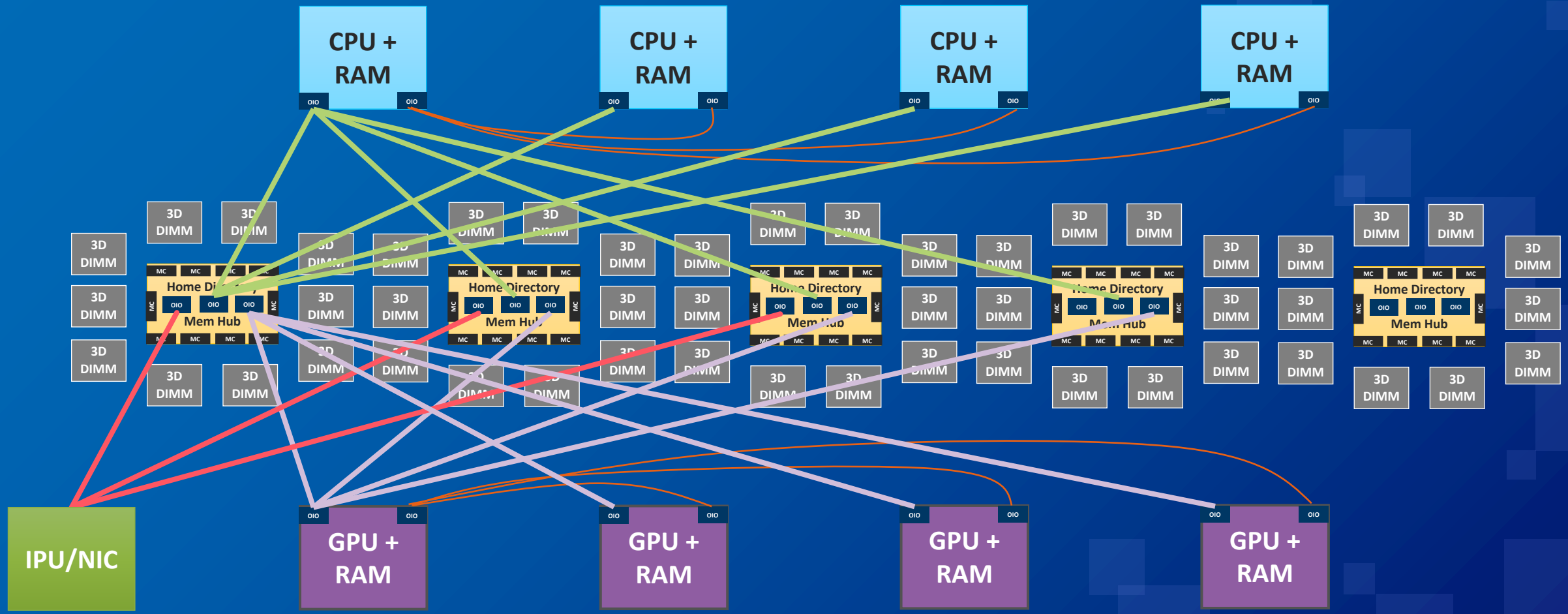
Capacity:BW scaling is off by >6x

Customers buy for bandwidth, but pay for (unused) capacity – creating a TCO problem at a global scale

Volume customers drive market solutions, dragging everyone along for the ride – whether or not they have the same problems

Re-thinking system architectures for Total Cost of Ownership

Correct pain by “right sizing” local memory and centralizing pooled resources for pathological corners

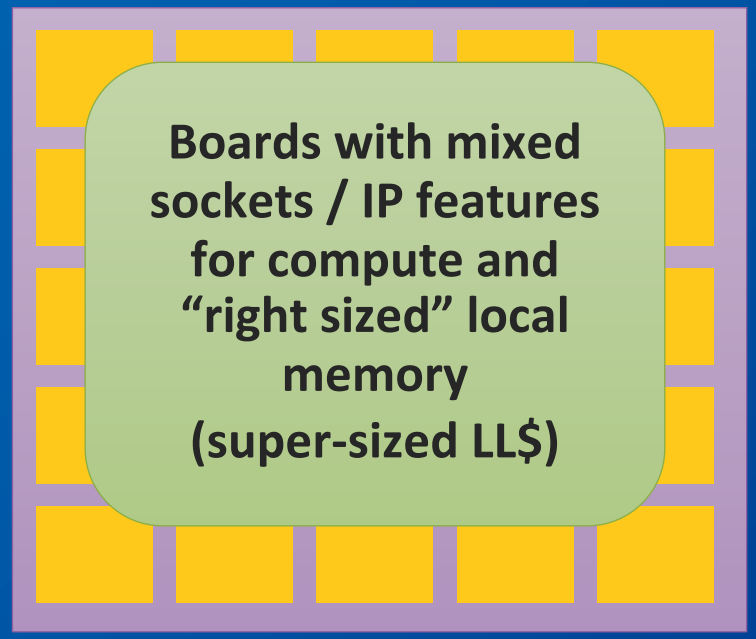


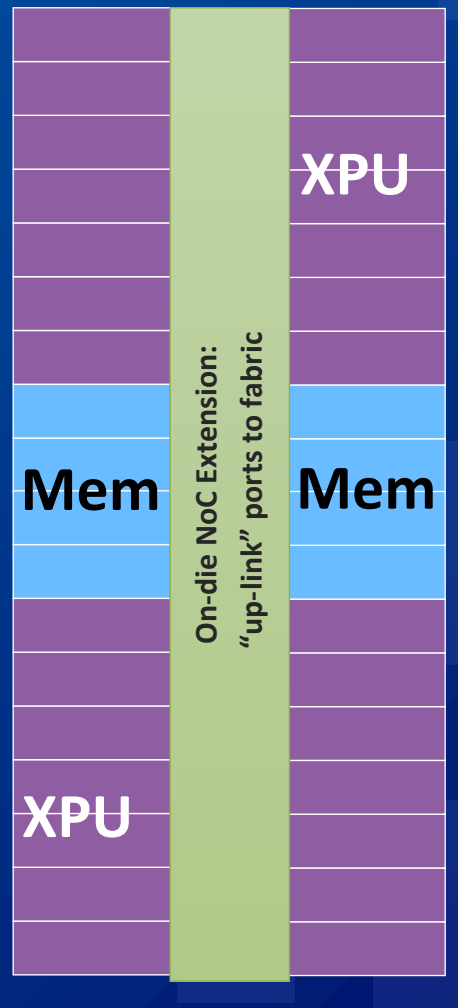
Not all links are drawn for simplicity

Is the future rack acting like today's socket?

xPU

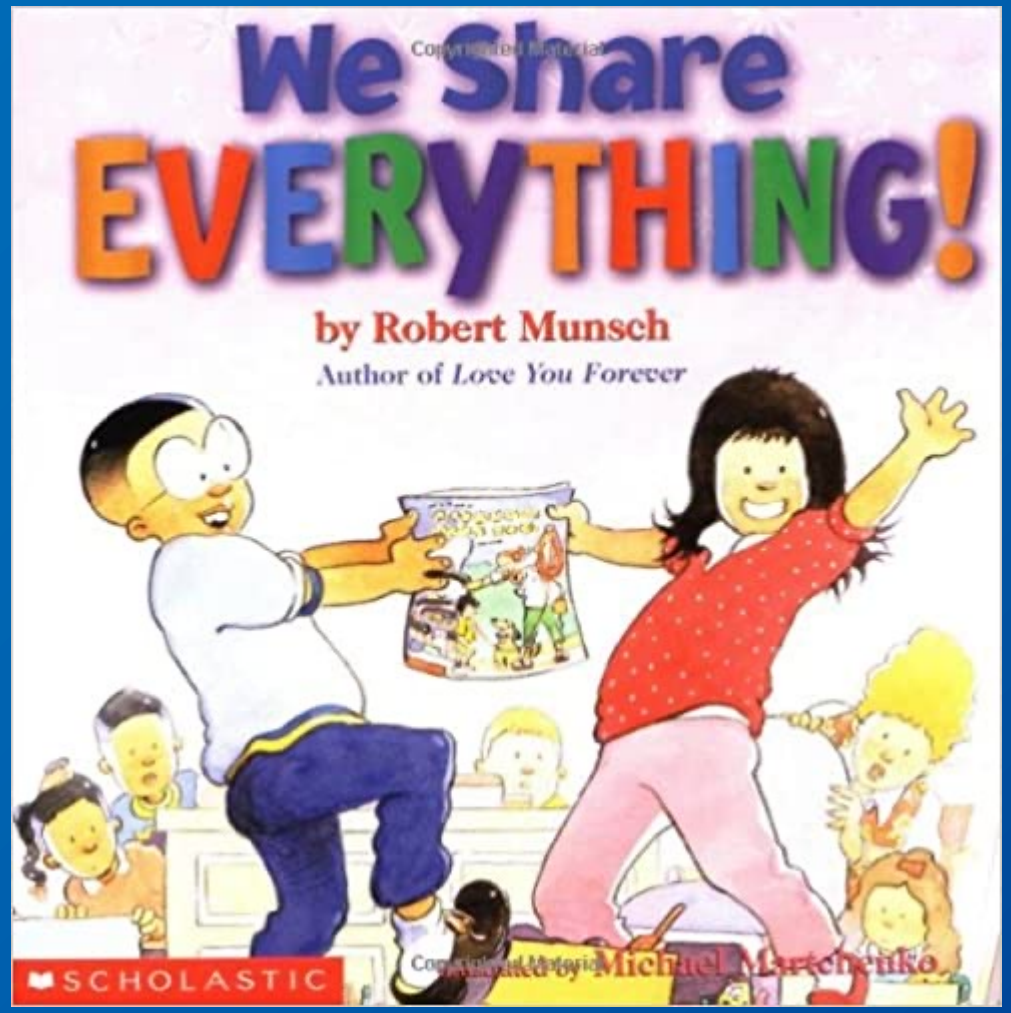
- OIO extends NoC
- Not old protocols
- Coherent
- Shared Addresses



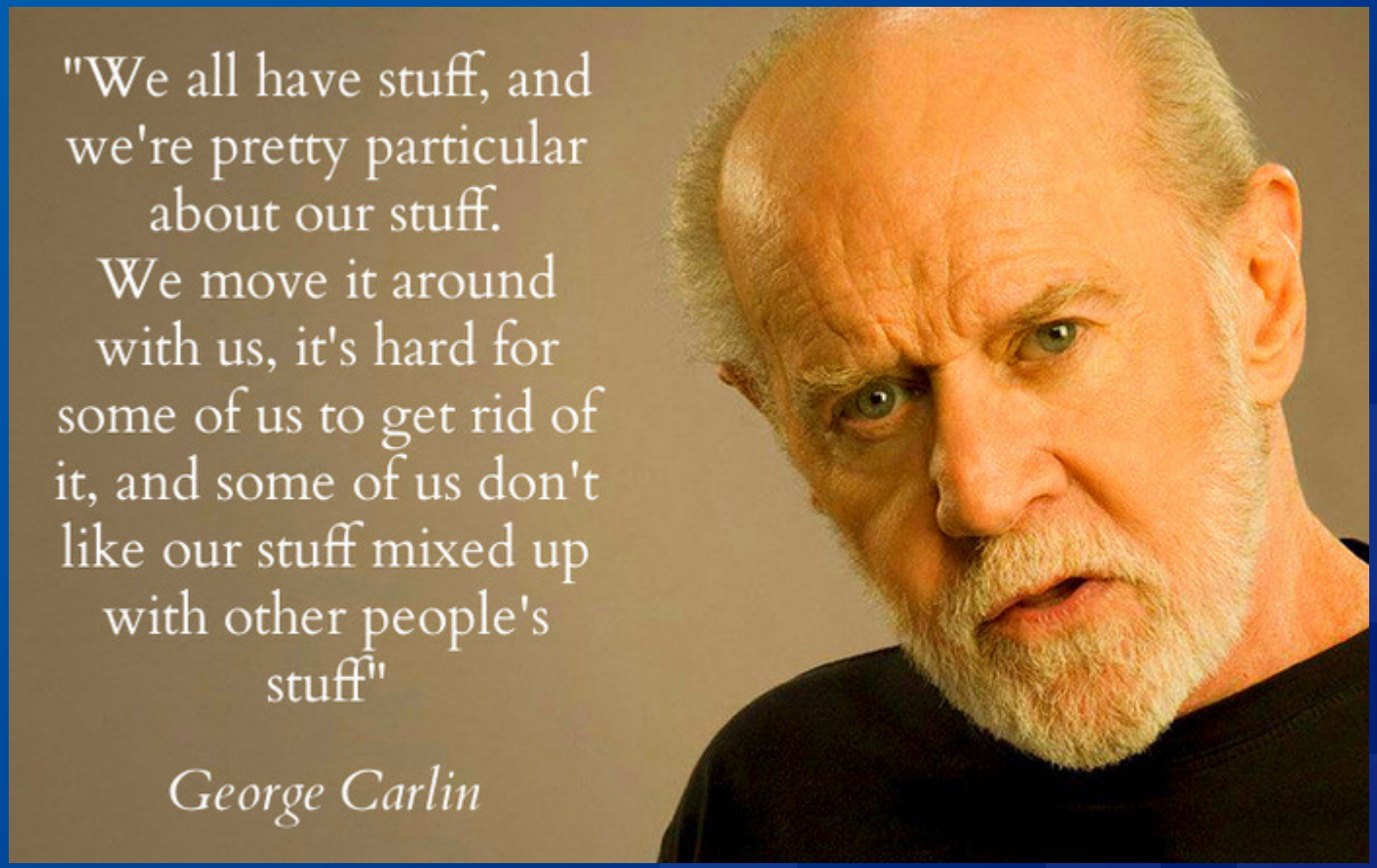


- Build "Mesh Extension" to every socket in the rack
- The "rack" becomes the logical "socket" boundary
- Shared address space enables "pointer passing" productivity

Two (Humorous) Views of This Model . . .

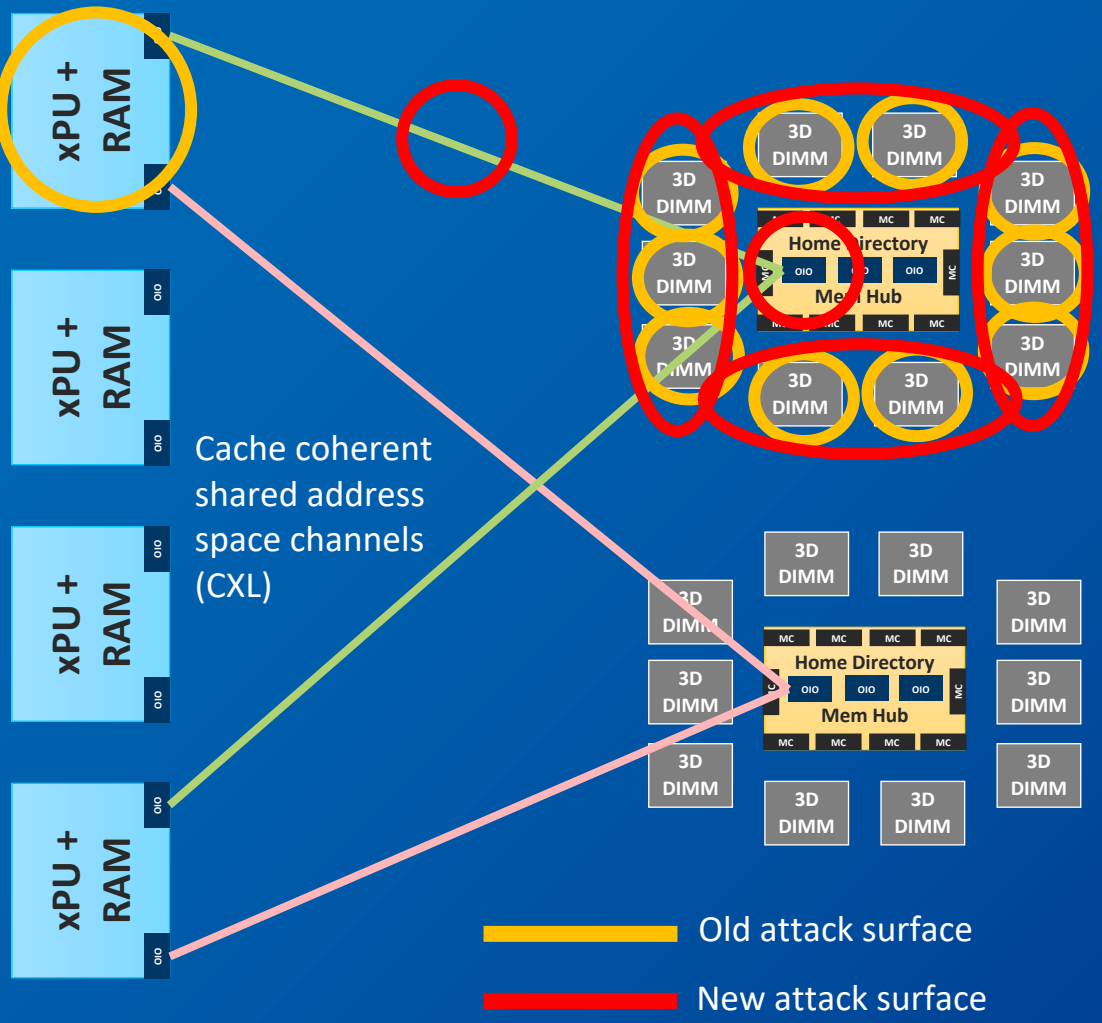


Source: https://www.amazon.com/We-Share-Everything-Robert-Munsch/dp/0590896016



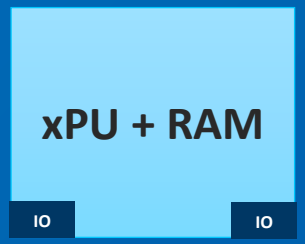
Source: https://www.deviantart.com/jev12345/art/George-Carlin-s-Stuff-Quote-920282794

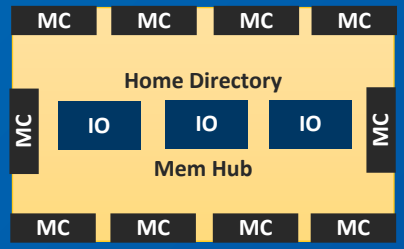
A closer look at “shared everything” platforms

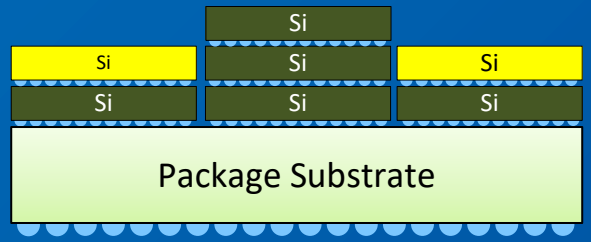


1. **Physical Separation?**
   - Replicate hardware: TCO problem!
2. **Temporal Separation?**
   - SLA/QOS terms: TCO problem!
3. **Logical Separation?**
   - Skip sharing via “ceiling” allocations: TCO problem!
4. **FHE Cryptographic Separation?**
   - Overheads (FLOPs, Bytes): TCO problem!
5. **New Attack Opportunities**
   - More side channels (exponential)
   - OS/MemHub resource handshake
   - Resource isolation
   - Malicious 3rd party IP “snooping”
   - **Blast radius of compromise**

This is also a “micro” problem – not just “macro”







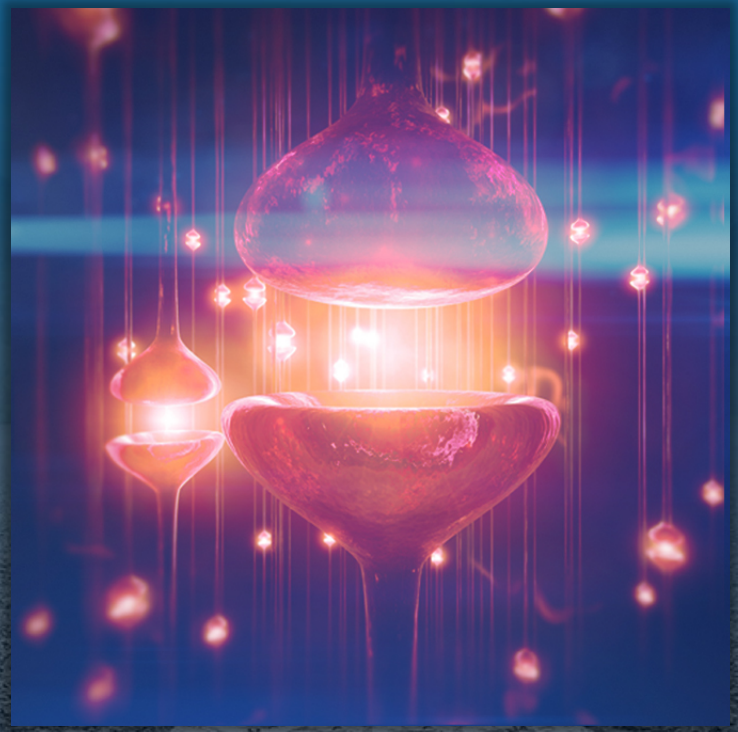
Cache-coherent shared address space interface for die-to-die operation (UCIe / CXL)

1. **Physical Separation?**
   - Replicate hardware: TCO problem!
2. **Temporal Separation?**
   - SLA/QOS terms: TCO problem!
3. **Logical Separation?**
   - Skip sharing via “ceiling” allocations: TCO problem!
4. **FHE Cryptographic Separation?**
   - Overheads (FLOPs, Bytes): TCO problem!
5. **New Attack Opportunities**
   - More side channels (exponential)
   - OS/MemHub resource handshake
   - Resource isolation
   - Malicious 3rd party IP “snooping”
   - **Blast radius of compromise**

Adapting to major industry trends

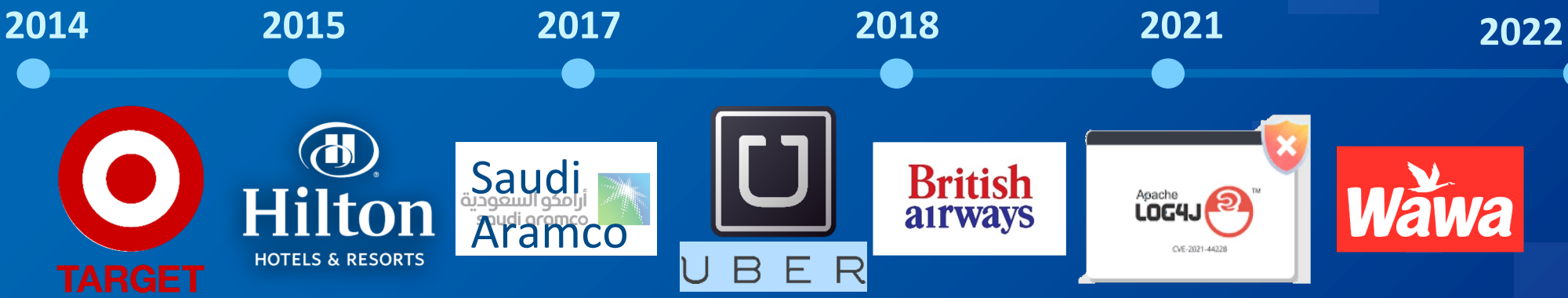
PROLIFERATION OF
Cloud Computing

GROWTH OF
AI & Analytics



CLOUDIFICATION OF THE
Network & Edge

Loss of IP is devastating



SQL injection

Privilege escalation

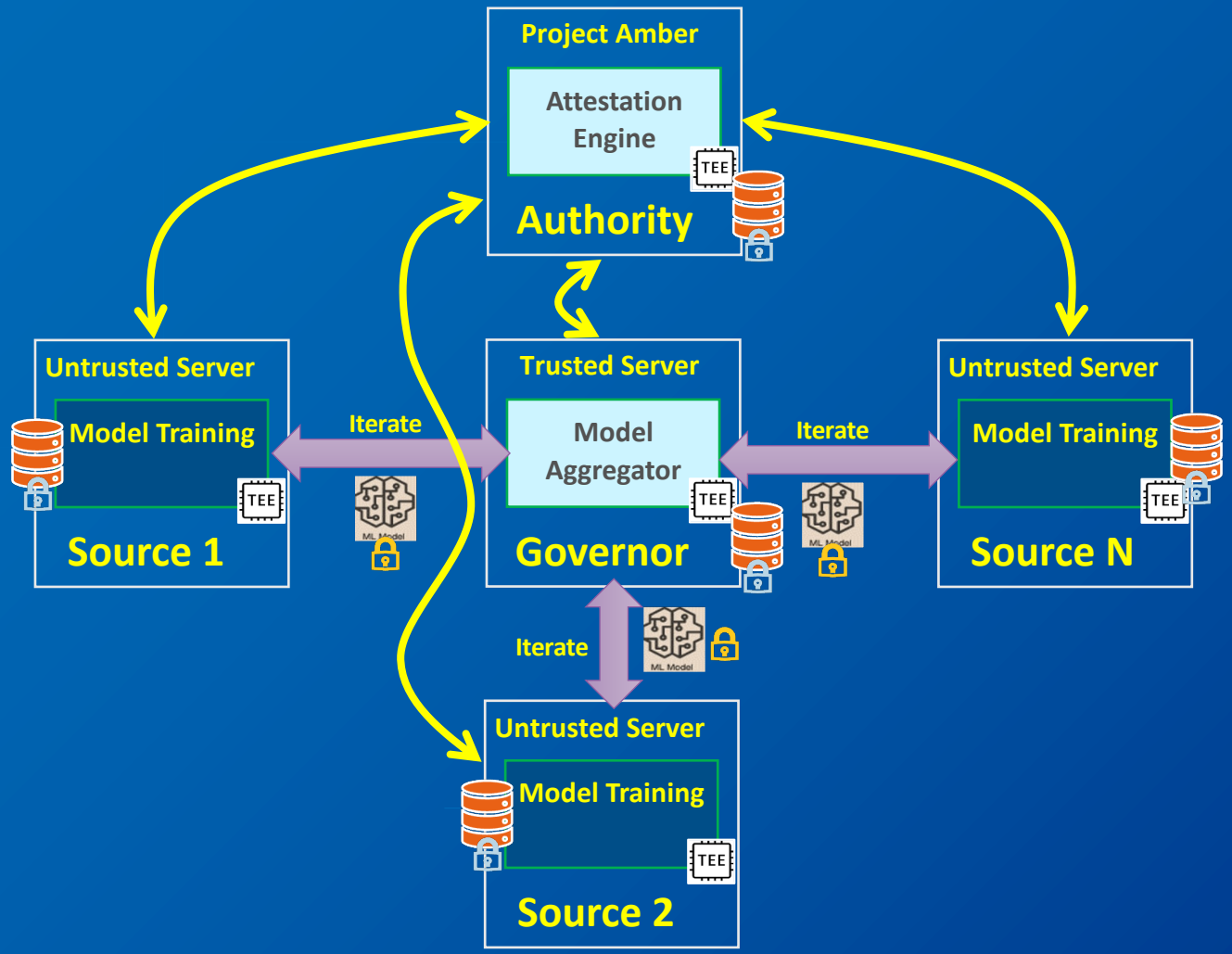
Memory scraping

$18 M	$10.5 T	$1.25 B	338 billion
Avg. cost of cybercrime per company	Global cybercrime annual costs by 2025	GDPR data breaches fines since Jan. 28, 2021	New Lines of new software code in 2025

Source: https://cybersecurityventures.com/the-world-will-store-200-zettabytes-of-data-by-2025/

Near-term Trust Example with Distributed AI

Secure Federated Learning with Confidential Computing and Project Amber

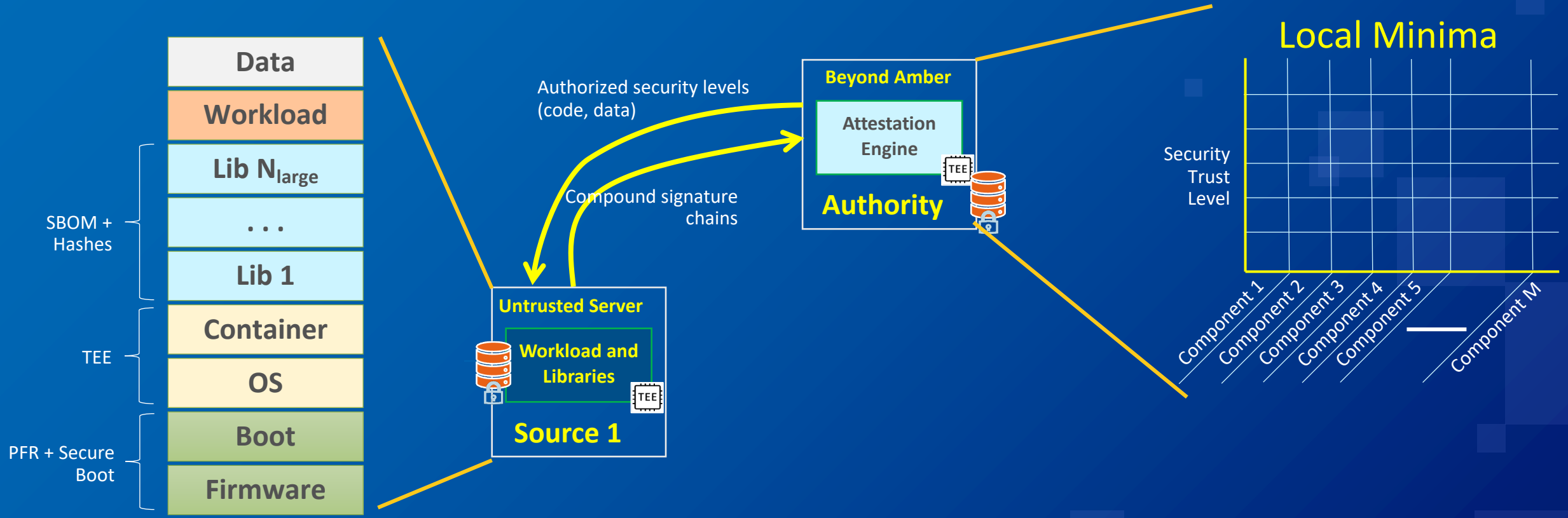


1. **AI Models Only Execute at Data Source**
   - Model close to data sources for training
   - Local training in TEE to protect privacy and IP
   - Infrastructure is *not* trusted
   - Actual data is never sent elsewhere
2. **Project Amber Attestation Service**
   - Project Amber verifies TEE trust worthiness
   - Each TEE infrastructure is verified separately
   - Confirm TEE trusted where model is trained
   - Compliance and Regulatory Validation
3. **Governor**
   - Governor collects models, not source data
   - Model updated and iterated back to sources
   - Repeated until model converges
   - Final model deployed for inference

Data is never unencrypted in transit, at rest, and in use. Model is protected by TEE during training.

Going Further: Whole Environment Attestation

Dynamic certification to handle *types* of data and code as conditions evolve – example: medical records



Validate entire HW/SW combination and recertify when any element changes

Approve algorithms (code) and data across a matrix of regulations and signatures

The importance of Attestation

Brittle measurements and implied properties

- Will this scale or will we be overwhelmed by complexity?
- Impact to workload performance: startup and interaction times?

How will we reason about…

- Inter-TEE security properties – fundamentally different TEEs
- Intra-TEE security properties – same TEE with different properties
  - E.g., evolution of building block technologies like TME-MK (IA) and Physical Memory Protection (RISC-V)
- TEE boundaries and SBOM coming together – signature explosion
- Policies of computers, nation-states, and technical domains


intel