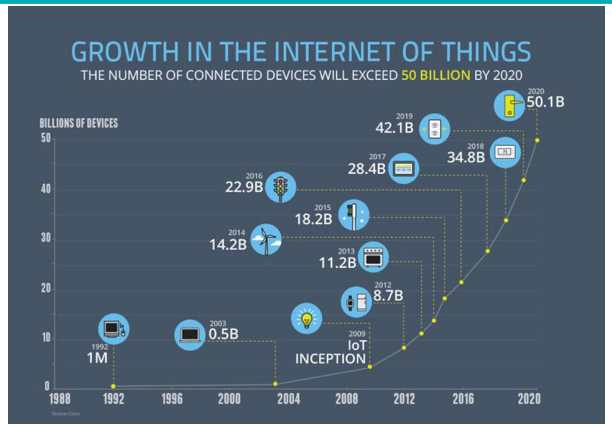
Gen-Z Technology: Enabling Memory Centric Architecture

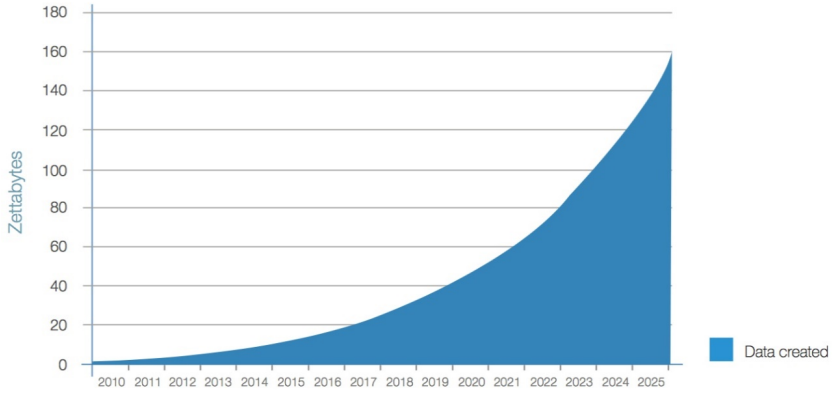
Kurtis Bowman
President of the Gen-Z Consortium
Director of Technologies & Concepts at DellEMC, Server Office of the CTO

Challenges In The Data Center

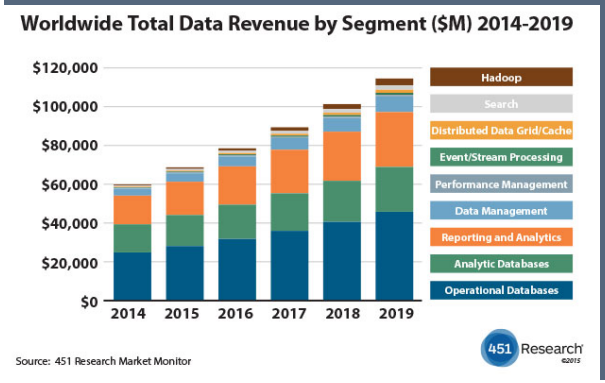
- More data than ever
  - 50 billion IoT devices by 2020
  - 4TB of data per day per autonomous vehicle
  - 160ZB of data a year by 2025
- CPU performance improvements have slowed
  - From 50% generation to generation to only 20%
- Managing resources
  - Key resources are trapped in today's servers
  - Resources are stranded when a failure occurs
- Maintaining a secure environment
  - The global average cost of a data breach is $3.86 million

Why Gen-Z?





Businesses' Need to 'Monetize' Data



Big Data

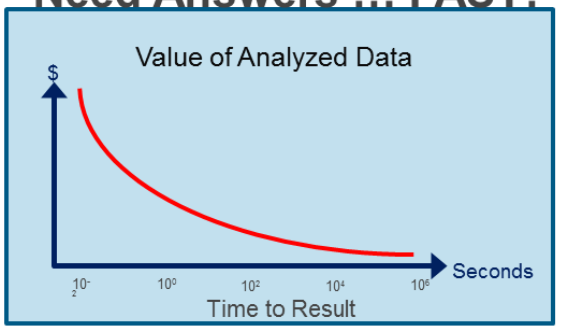
AI

Machine Learning

Deep Learning

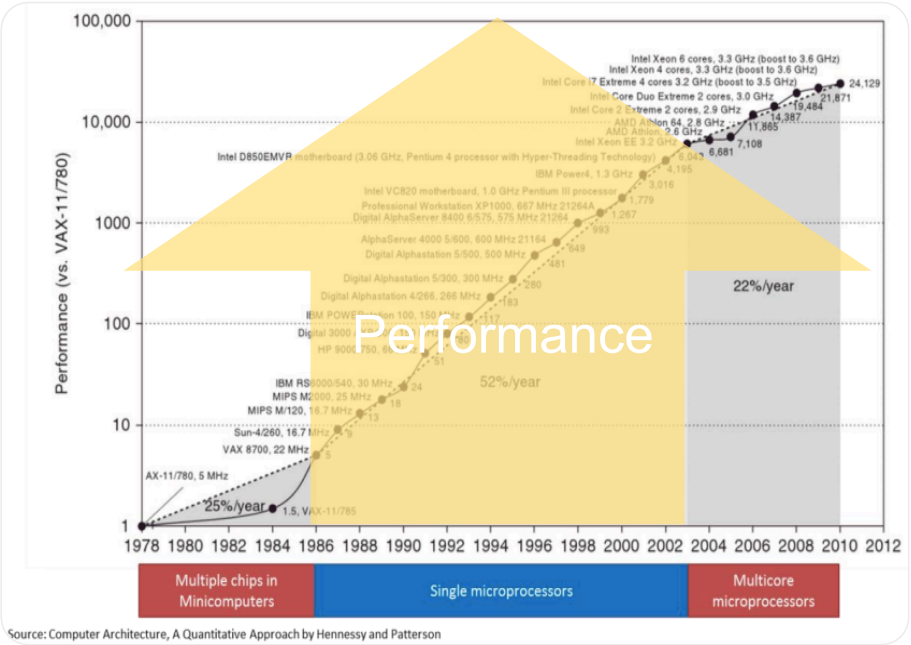
BI

Need Answers … FAST!

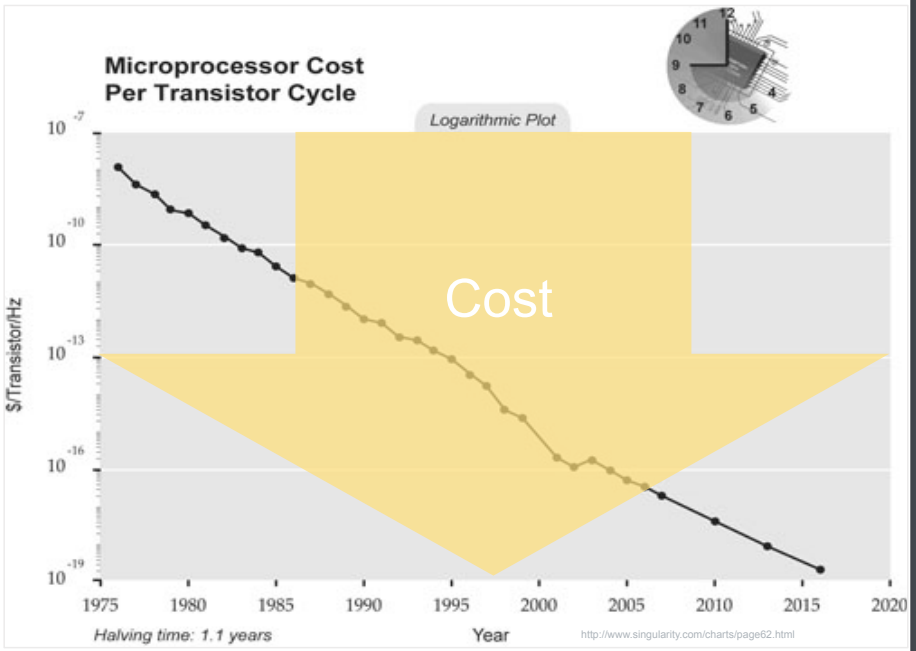


- Businesses demanding real-time insight
- Increasing amounts of data to be analyzed

CPU Performance Improvements Have Stalled



Source: Computer Architecture, A Quantitative Approach by Hennessy and Patterson



Driving new Architectures

Where & how we spend transistors is changing

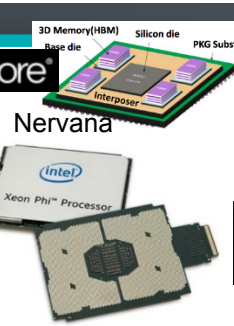
Date Production Part Available*

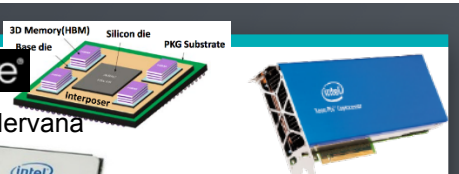
Foundry	2012	2013	2014	2015	2016	2017	2018	2019	2020
tsmc	28HPM		20SoC 28HPC	16FF+	16FFC	10FF	7FF	7HPC	5nm
SAMSUNG			20LPE	14LPE	14LPP	10LPE	10LPP	7nm	
GLOBALFOUNDRIES					14LPP	22FDX		7nm 12FDX	
intel	22nm	22SoC	14nm	14SoC	14nm+	10nm 10SoC			7nm

*risk production and qualification start is typically 1 year ahead

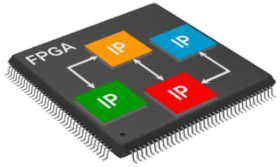
Tech Insights

Specialty Processors



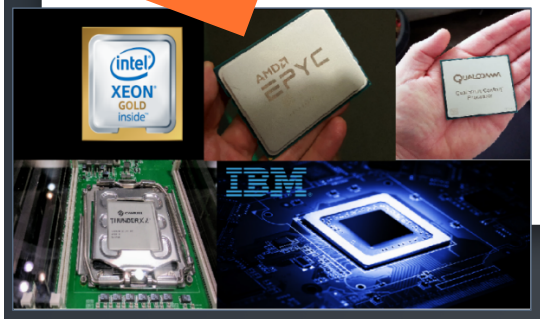


FPGAs

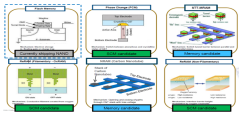


GPUs

CPU Competition re-Emerges



Storage Class Memory

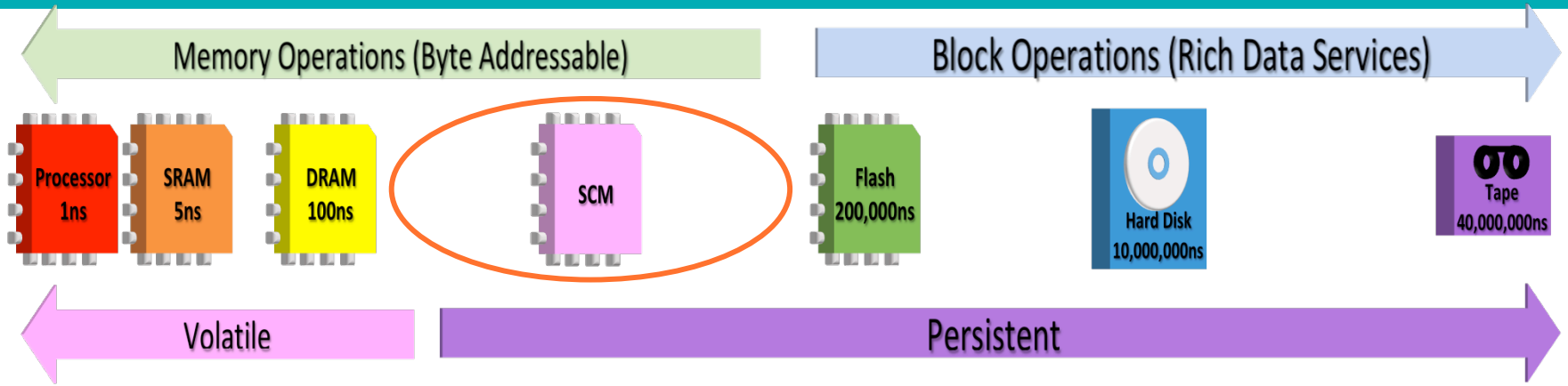


SCM: PCM, 3DxPt, STT-MRAM, ReRAM, NRAM

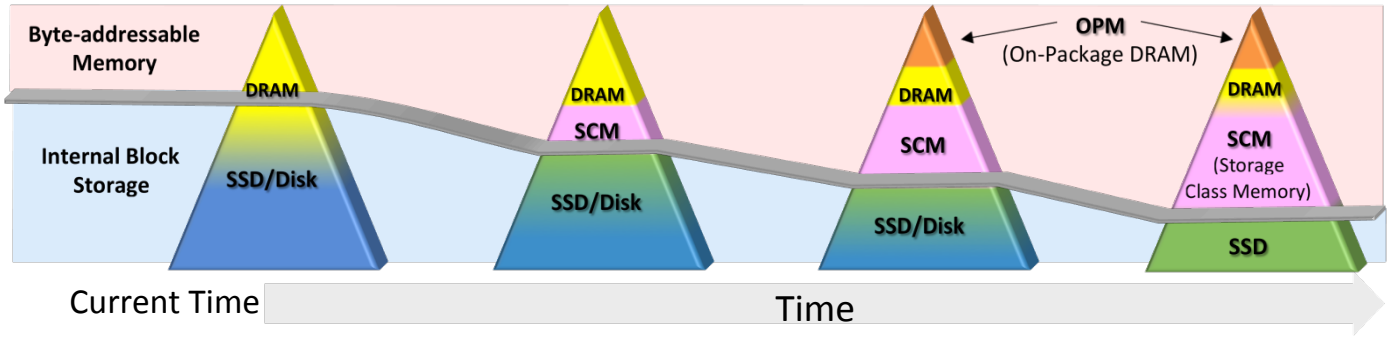
NVDIMM-N/P

SmartNICs

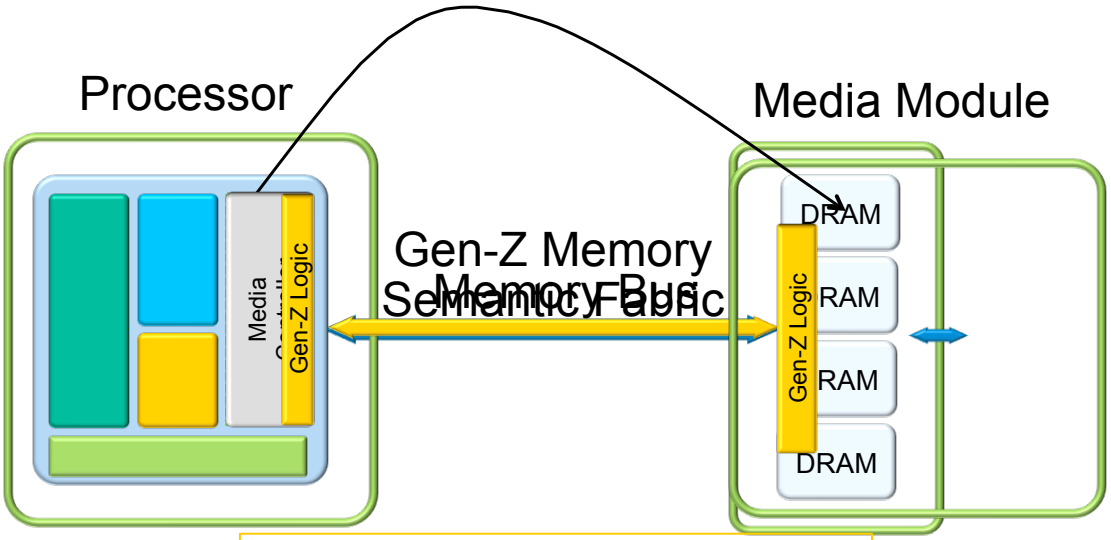
Memory and Storage are Converging



With memory/storage convergence, memory semantic operations become predominant (volatile & non-volatile)



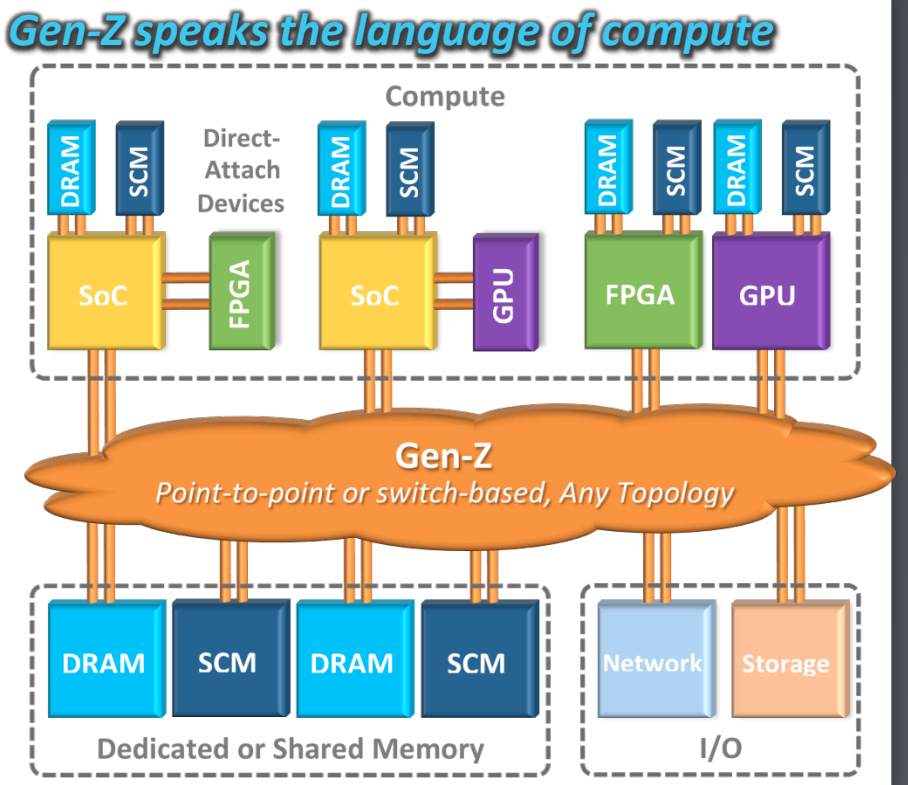
Gen-Z Allows Memory Innovation



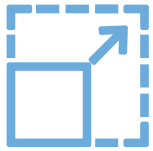
2-8 High-speed Serial Links
4-8 Memory Channels
Low Latency, High performance
17-25 GB/s / Channel
Split Memory Controller
288pins / DIMM
Asynchronous Interface
Synchronous Interface
Processor is media agnostic

Gen-Z Connects Disaggregated Components

- High Performance
  - High Bandwidth, Low Latency, Scalable
  - Eliminates protocol translation cost / complexity / latency
  - Eliminates software complexity / overhead / latency
- Reliable
  - No stranded resources or single-point-of-failures
  - Transparently bypass path and component failure
  - Enables highly-resilient data (e.g., RAID / erasure codes)
- Secure
  - Provides strong hardware-enforced isolation and security
- Flexible
  - Multiple topologies, component types, etc.
  - Supports multiple use cases using simple to robust designs
  - Thorough yet easily extensible architecture
- Compatible
  - Use existing physical layers, no OS modifications required
- Economic
  - Lowers CAPEX / OPEX, unlocks / accelerates innovation



Disaggregated infrastructure benefits



Compose servers with resources app requires

Unlock trapped resources

Avoid overprovisioning

Purchase resources independently

INCREASE AGILITY **OPERATE EFFICIENTLY** **UNLOCK VALUE**

Increase RAS

Repurpose retired resources

Technologies can evolve – and deployed independently

Security Is Part Of The Gen-Z Fabric

- In today's environment every device is a potential threat
  - Inflight attacks, denial of service attacks, packet injection attacks, time manipulation attacks, and more
- Gen-Z has the tools needed to handle these threats
  - Data encryption and cryptographically-secure message authentication
  - Tight timeout domains with immediate response scheduling
  - Access Keys (A-Key) component group-level access control
  - Region Keys (R-Key) page level access control
  - Replay attack detection
  - Packet destruction detection
  - Extreme packet injection rate protection
  - Data destruction protection
  - Resource exhaustion protection

Open Consortium With Broad Industry Support