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!

Value of Analyzed Data

- Businesses demanding real-time insight
- Increasing amounts of data to be analyzed
CPU Performance Improvements Have Stalled

![Graph showing the decline in CPU performance improvements and the increase in cost over time.]

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

Driving new Architectures
Where & how we spend transistors is changing

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>TSMC</td>
<td>25HP</td>
<td>20nm</td>
<td>14nm</td>
<td>10nm</td>
<td>7nm</td>
<td>7nm</td>
<td>7nm</td>
<td>7nm</td>
<td>7nm</td>
</tr>
<tr>
<td>SAMSUNG</td>
<td>20LP</td>
<td>14LP</td>
<td>10LP</td>
<td>7nm</td>
<td>7nm</td>
<td>7nm</td>
<td>7nm</td>
<td>7nm</td>
<td>7nm</td>
</tr>
<tr>
<td>AMD</td>
<td>28nm</td>
<td>22nm</td>
<td>14nm</td>
<td>10nm</td>
<td>7nm</td>
<td>7nm</td>
<td>7nm</td>
<td>7nm</td>
<td>7nm</td>
</tr>
<tr>
<td>INTEL</td>
<td>32nm</td>
<td>22nm</td>
<td>14nm</td>
<td>10nm</td>
<td>7nm</td>
<td>7nm</td>
<td>7nm</td>
<td>7nm</td>
<td>7nm</td>
</tr>
</tbody>
</table>

*Note: Production and qualification start is typically 1 year ahead.

CPU Competition

Storage Class Memory

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

SmartNICs

FPGAs

Specialty Processors

GPU
Memory and Storage are Converging

With memory/storage convergence, memory semantic operations become predominant (volatile & non-volatile)
Gen-Z Allows Memory Innovation

- Processor
  - 4-8 Memory Channels
  - 17-25 GB/s / Channel
  - 288 pins / DIMM
  - Synchronous Interface
  - Gen-Z Logic
  - Media Module
    - DRAM
    - RAM
    - Asynchronous Interface
    - Low Latency, High-performance
  - Split Memory Controller
  - 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**
- Increase RAS

**OPERATE EFFICIENTLY**
- Repurpose retired resources

**UNLOCK VALUE**
- 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

#### Consortium Members

| Allion Labs | IDT |
| Alpha Data | IntelliProp |
| AMD | Jess Link |
| Amphenol | Keysight |
| ARM | Lenovo |
| Avery Design Systems | Lotes |
| Broadcom | Luxshare-ICT |
| Cadence | Mellanox |
| Cavium | Micron |
| Cisco | Microsemi |
| Cray | Mobiveil |
| Dell EMC | Molex |
| Everspin | NetApp |
| ETRI | Nokia |
| FIT | Oak Ridge Natl Labs |
| Google | PLDA Group |
| Hirose | Qualcomm |
| HPE | Red Hat |
| Huawei | Samsung |
| IBM | Seagate |
| Senko Advanced Comp | Simula Research Lab |
| SK hynix | Smart Modular |
| Smart Modular | Spin Transfer Tech |
| Spin Transfer Tech | Teledyne LeCroy |
| Toshiba Memory Corp | TE |
| Univ. New Hampshire | VMware |
| Western Digital | Xilinx |
| Yadro | Yonsei University |

*Board member
*Associate member