High-Perf Array Optimization

Presenter: Odie Killen, VP of Engineering at Viking Enterprise Solutions

FMS the Future of Memory and Storage

Overview

- Capacity, Performance or Both?
- Capacity Capabilities and Challenges
- Performance Challenges – SW
- Performance Challenges – HW
- Ways to Solve these Challenges
- Unbalanced Topologies
- Typical Balanced Topology
- Example – Gen4/Gen5
- Example – Unbalanced Gen5
- Path Forward

Capacity, Performance or Both?

- Modern SSDs create a problem, design for capacity, performance or both?
- U.2 SSDs are shipping up to 60TB today, with larger capacities coming
  Large capacity means large failure blast radius
- Single socket CPU instances typically allocated 64 lanes for SSD connectivity
  252 GB/sec (64x PCIe Gen5 Lanes for host and SSDs)
- PCIe Gen5 capable SSDs can support 14 Gb/sec SQ RD and 8.6 GB/sec SQ WR
  x4 lanes connection, assuming 128 lanes => 448 GB/sec SQ RD, 275 GB/sec SQ WR
- Typical 19” chassis can support large numbers of SSDs
  24 – 80+ SSDs depending on chassis configuration
- Current device capabilities drive many trade offs in system design

Capacity Capabilities and Challenges

- U.2 SSDs are shipping up to 60TB today, with larger capacities coming
  - Large capacity means large failure blast radius
  - Longer rebuild times and increased CPU consumption to recover from failures
  - Longer time operating in a degraded state
- Typical 19” chassis can support large numbers of SSDs
  - 24 – 80+ SSDs depending on chassis configuration
  - Can easily deploy more SSDs than available PCIe lane connections
  - Typical HA solution, 2 lanes per controller to each SSD
  - More than 32 drives and you have created a bottleneck
- Mechanical solutions can be achieved that support more SSDs than can be used from a performance perspective
  - Performance is bottlenecked due to device density
  - Additional cost is driven with no additional performance improvement

Performance Challenges – SW

- Balanced HW solutions (ingress = egress lane counts) are inherently inefficient
- SW is a bottleneck, HW design can support a completely balanced solution
- SW processing of data creates bottlenecks
  - Very efficient solutions are 70% efficient (70% of line rate)
  - NICs are 85-90% efficient, OS is 80% efficient
  - With 64x PCIe Gen5 lanes in, expect no more than 176 GB/sec per controller
- With SW overhead, it is not possible to process data at line rates
  - CPU cycles, memory hops and other processing steps consume cycles and create a reduction from line rate speeds
- Even best in class SW is not capable of leveraging a balanced HW topology

Performance Challenges – HW

- General purpose storage devices (as built by most OEM/ODMs) employ balanced topologies
  - BW is nearly always not 100% consumed in this topology
  - PCIe Gen5 speeds drive higher design costs due to material and placement requirements
  - General purpose topologies are not optimized for SW solutions
  - Can design with more devices than can be effective used
- Purpose built solutions are not easily optimized for general applications
- Gen5 SI requirements dictate high cost material and distributed switching
  - High speed interface has higher signal loss requiring better material
  - Signal losses dictate retimers or switches
- For solutions that can’t be direct connected to CPU, switch fabrics are required
  - High device solutions will mandate distributed switching

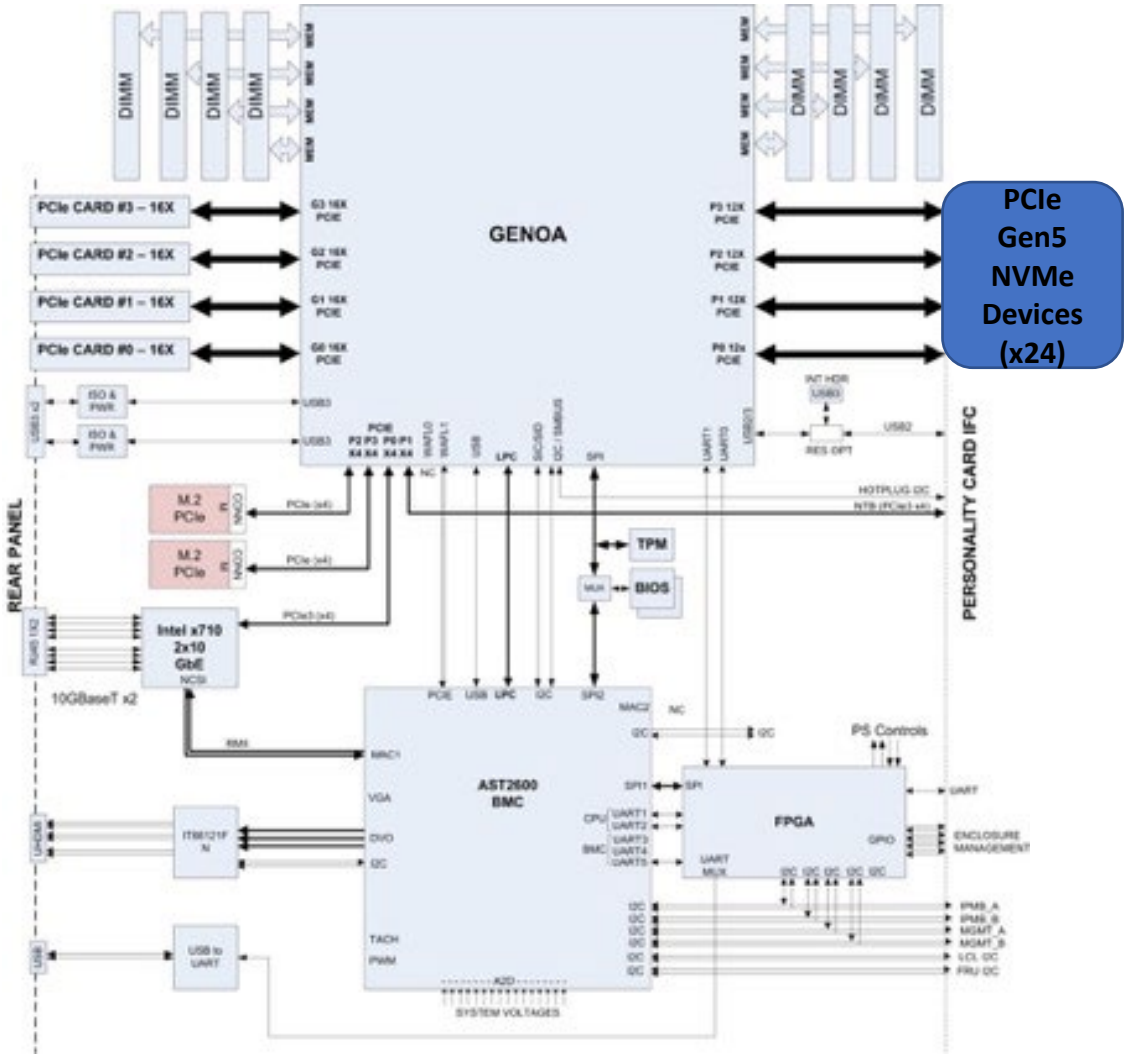
Ways to Solve These Challenges

- Make the ingress bottleneck BIGGER
- Change the traditional approach of architecting high performance arrays
- Design HW to be more customized for specific SW stacks
  - Less general purpose architectures which are typically balanced
- Account for all possible BW implications in design and optimize based on data
- Implement un-balanced architectures that provide more front end BW
  - Back end BW is limited by ingress data and processing
  - Unbalanced architectures allow greater ingress data BW
- Design for either High Performance or for Capacity

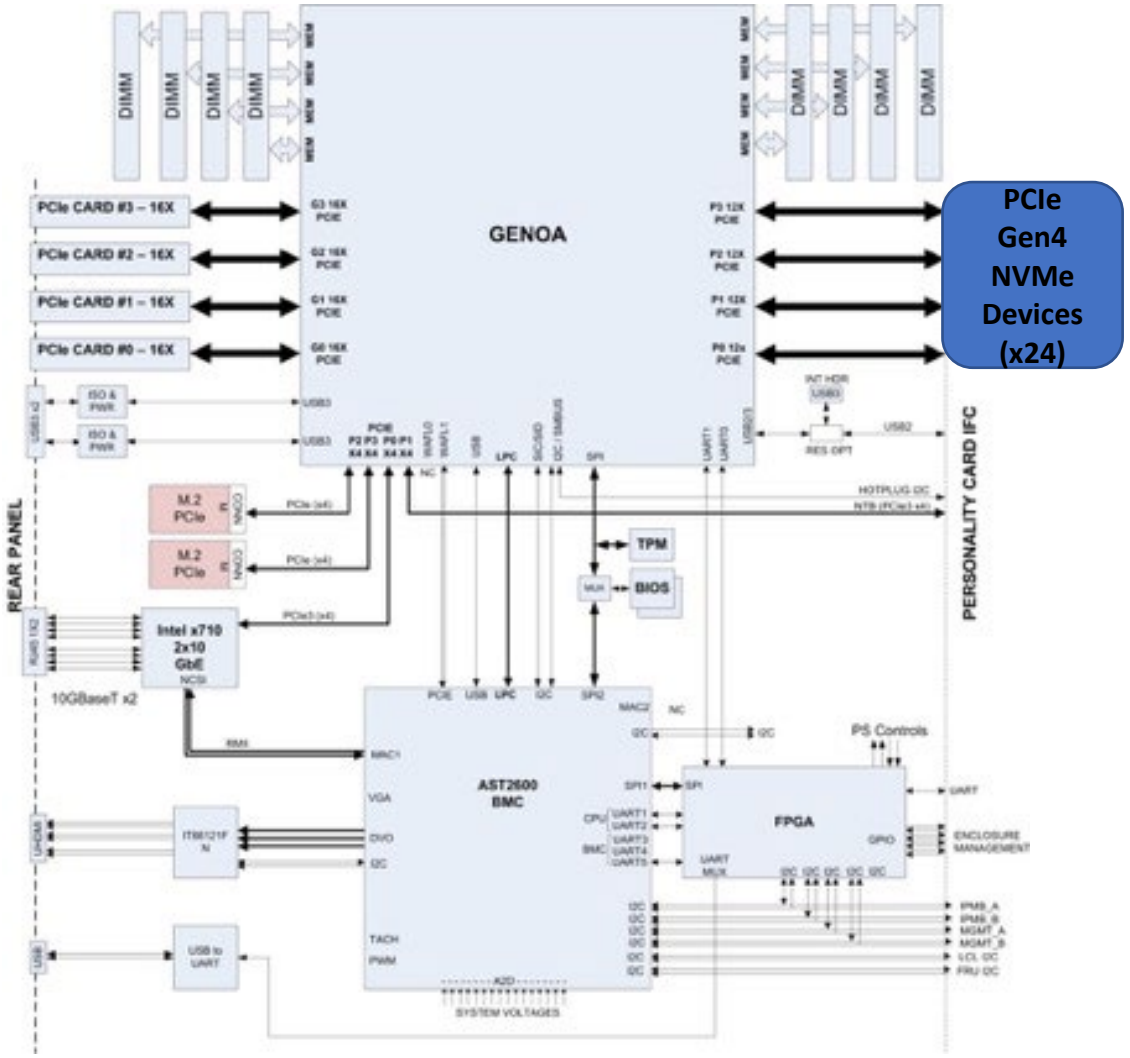
Unbalanced Topologies

- Unbalanced topologies present non-traditional solutions to optimize performance
- Gen5 front ends with Gen4 back ends
  - In a balanced lane topology, a slower back end can help account for SW inefficiency
  - Gen4 back end is lower cost and less complicated to implement
  - Less efficient SW stacks can be 50-60% efficient, thereby optimized for a solution like this
  - Lower cost Gen4 SSDs
- Larger number of host side connections versus device side connections
  - Typical CPUs offer 128 lanes of Gen5 connectivity
  - Implement 80 host lanes (5 x16) and 48 (24 x2) back end device lanes
  - Results in a topology optimized for 60% efficient SW/CPU overhead
  - Good balance of HW optimized for SW implementation
  - A higher percentage of available BW will be consumed versus balanced approaches

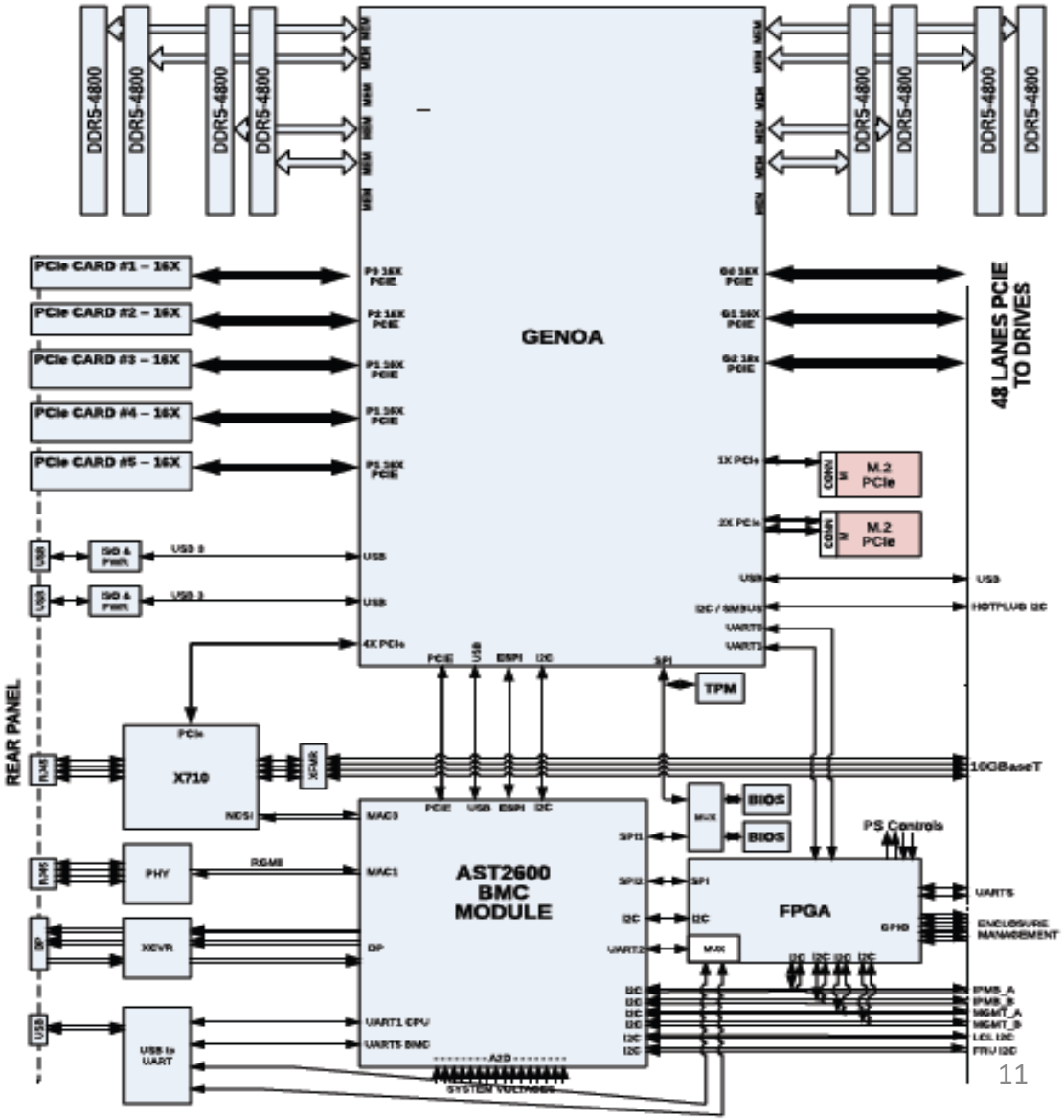
Typical Balanced Topology



Example – Gen4/Gen5



Example – Unbalanced Gen5



Path Forward

- More purpose built architectures
  - Unbalanced approaches to optimize BW usage
- Design for high performance or capacity
  - Don’t try to make one solution work for all applications
- Leverage more lower cost Gen4 based SSDs
- Design topology to fully leverage available performance from higher cost Gen5 devices