NVIDIA GRACE
Datacenter Ready

- NVIDIA's First Server CPU
- 72 Arm v9.0 cores
  - SVE2 support
  - Virtualization Extensions: Nested Virtualization, S-EL2 support
- RAS v1.1
- GIC v4.1
- SMMU v3.1

- Built on TSMC 4N process node
NVIDIA GRACE
Designed from the Ground-Up to be a Superchip
**NVLINK-C2C**
High Speed Chip to Chip Interconnect

- Used to create the Grace Hopper, and Grace Superchips
- Removes the typical cross-socket bottlenecks
- Up to 900GB/s of raw bidirectional BW
  - Same BW as GPU to GPU NVLINK on Hopper
- Low power interface - 1.3 pJ/bit
  - More than 5x more power efficient than PCIe
- Enables coherency for both Grace and Grace Hopper superchips
GRACE SUPERCHIP
Standards Compliant Platform

- Targets Arm standards for off the shelf OS compatibility
- Arm Server Base System Architecture (SBSA)
- Arm Server Base Boot Requirements (SBBR)
- Arm Memory Partitioning and Monitoring (MPAM)
- Arm Performance Monitoring Units (PMUs)
GRACE HOPPER
Heterogenous Coherency

- Unified Memory with shared page tables
  - Shared CPU and GPU virtual address space
  - GPU access to pageable memory
  - System allocator support for GPU memory
    - Yes, malloced and mmaped pointers!
- Native atomics, including standard C++ atomic support
Enables remote NVLINK connected GPUs, to access Grace’s memory at native NVLINK speeds
NVIDIA GRACE
NVIDIA Scalable Coherency Fabric

- NVIDIA fabric and distributed cache design
- 3,225.6 GB/s Bi-section BW
- Scalable to 72+ cores
- 117MB of L3 cache
- Arm Memory Partitioning and Monitoring (MPAM)
- Supports up to 4-socket coherency over Coherent NVLINK

*Example possible fabric topology for illustrative purposes*
NVIDIA GRACE — SCF
Memory Partitioning and Monitoring

- Arm standard for partitioning system resources
- Partition IDs (PARTID) are assigned to entities making requests to memory
- SCF Cache resources can be partitioned between different PARTIDs
- Partition both Cache capacity, and Memory Bandwidth
- Performance Monitor Groups (PMG) can be used to monitor resource usage
NVIDIA GRACE
Memory

- Up to 512GB of LPDDR5x memory
- 32 channels
- Up to 546 GB/s of memory BW
- But why LPPDR?
NVIDIA GRACE
Remember the Superchips?

Grace is always paired
# MEMORY CHOICES
HBM, DDR, or LPDDR?

<table>
<thead>
<tr>
<th></th>
<th>HBM2e (4-sites)</th>
<th>DDR5 (8-channel)</th>
<th>LPDDR5x (32-channel)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacity</td>
<td>64GB</td>
<td>Up to 4TB</td>
<td>Up to 512GB</td>
</tr>
<tr>
<td>BW</td>
<td>Up to 1.8TB/s</td>
<td>Up to 358GB/s</td>
<td>Up to 546GB/s</td>
</tr>
<tr>
<td>Power/GBps</td>
<td>1x</td>
<td>8x</td>
<td>1x</td>
</tr>
<tr>
<td>Cost/GB</td>
<td>&gt;3x</td>
<td>1x</td>
<td>1x</td>
</tr>
</tbody>
</table>

Remember? C2C BW - 900 GB/s
NVIDIA GRACE
How Much Memory Do I Need?

- Natural Language Processing
- GPT-3 inference – fp8 – 175GB of memory
- GPT-3 training – over 2.5TB of memory

- Extended GPU Memory to the rescue!
- 4x decrease in the number of GPUs needed to fit the training set in memory
GRACE I/O

- Up to 68 lanes of PCIe
  - 4 — GEN5 x16 links
  - 128 GB/s bi-dir per x16
  - 2 — x2’s for misc
- Up to 12 lanes of coherent NVLINK
  - Shared with two — GEN5 PCIe x16
- NVLINK-C2C
  - 900 GB/s of raw bi-dir BW
NVIDIA GRACE SUPERCHIP
Purpose built for Supercomputing and HPC

- 144 Arm v9.0 cores with SVE2
  - Single thread perf optimized
- Up to 1TB/s memory bandwidth
- NVLINK-C2C for 3x typical inter-chip bandwidth
- Energy efficient design with LPPDDR5 allowing more power for compute
  - 500W TDP, core + memory
SPEC RATE — ESTIMATES

Source: Pre-silicon Estimated Performance results with GCC compiler, subject to change
GRACE CPU MEMORY BENCHMARK

Source: NVIDIA Grace pre-silicon results for single Grace SOC, subject to change
HOPPER GPU TO GRACE MEMORY BENCHMARK

Source: NVIDIA Grace Hopper pre-silicon results, subject to change
NVIDIA GRACE

Summary

- NVIDIA's First Server SOC
- 72 Arm v9.0 CPU cores
- NVLINK-C2C
- LPDDR5x for low power, high bandwidth
- Extended GPU memory (EGM) for scale out

Thanks to the entire Grace CPU team!