Hewlett Packard Enterprise

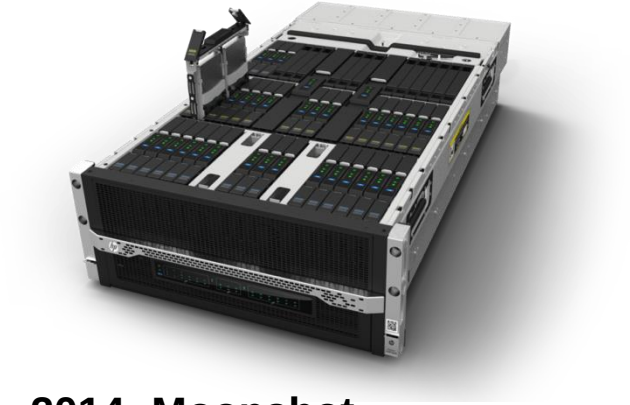
Update From the Coalface of Arm in HPC

Andy Warner, Distinguished Technologist

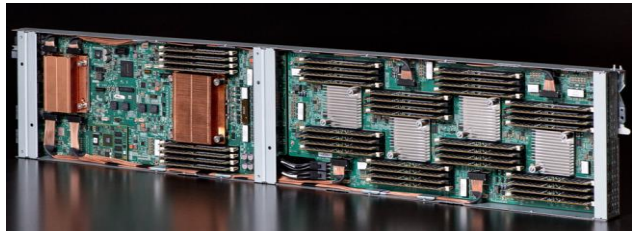
Arm HPC User Group | 20 June 2019 | Frankfurt, Germany

HPE and ARM: Long-standing Partnership

2012: Redstone
– Calxeda low power CPUs
– 288 nodes in 4U
– ARMv7 32bit
– 4 cores, 1.4GHz

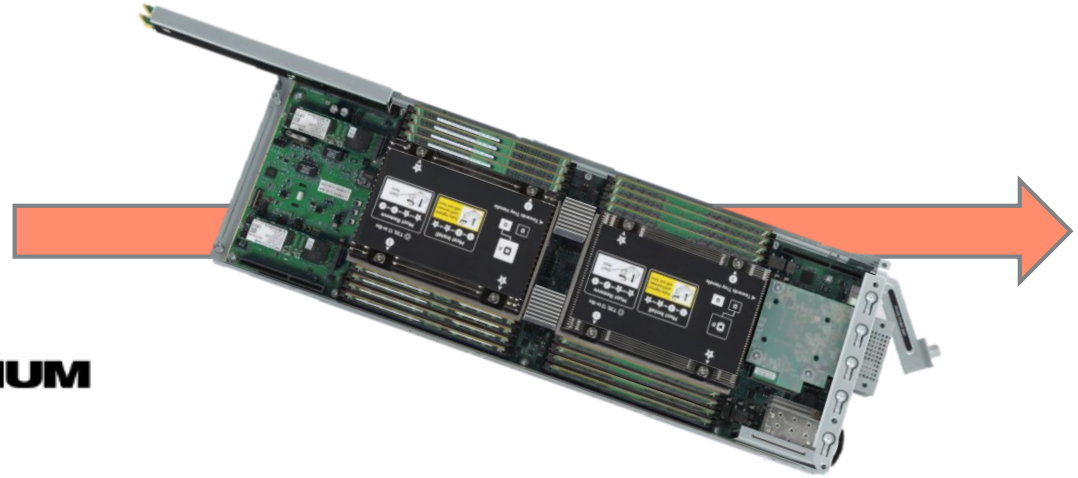


2014: Moonshot
– Calxeda, TI, Applied Micro
– 45 XGene cartridges in 4U
– ARMv8, 64 bit
– 8 cores, 2.4 GHz



2016: "The Machine" (prototype)
– Broadcom Vulcan CPU
– 160 TiB of addressable memory
– Gen-Z fabric
– Fabric Attached Memory
– Integrated fabric optics

2017: Commanche
– Cavium TX2 Early Access
– Four 2P nodes in 2U
– 32 cores, 2.2GHz



2018: Apollo70
– 28c & 32c SKUs offered
– Astra - Top500 system
– CatalystUK

Experience at Scale

There are other presentations at this workshop which will present detailed project or machine updates; I am not going to preempt them here. However, I believe the software ecosystem has proven remarkably – boring*.

Pushing systems at scale have uncovered remarkably few system software issues; but the ones that we have found are non-trivial, including:

```
[172322.408000] Unable to handle kernel paging request at virtual address ffffffffffffa0
[172322.416001] Mem abort info:
[…]
[172322.455200] Internal error: Oops: 96000005 [#1] SMP
[…]
[172322.746745] [<ffff0000082cd5cc>] dcache_readdir+0x9c/0x170
[172322.752306] [<ffff0000082b4fd8>] iterate_dir+0x150/0x1b8
[172322.757691] [<ffff0000082b5780>] SyS_getdents64+0x98/0x170
[…]
```

a.k.a. Redhat BZ 1702057 - Kernel panic on job cleanup, related to SyS_getdents64

and:

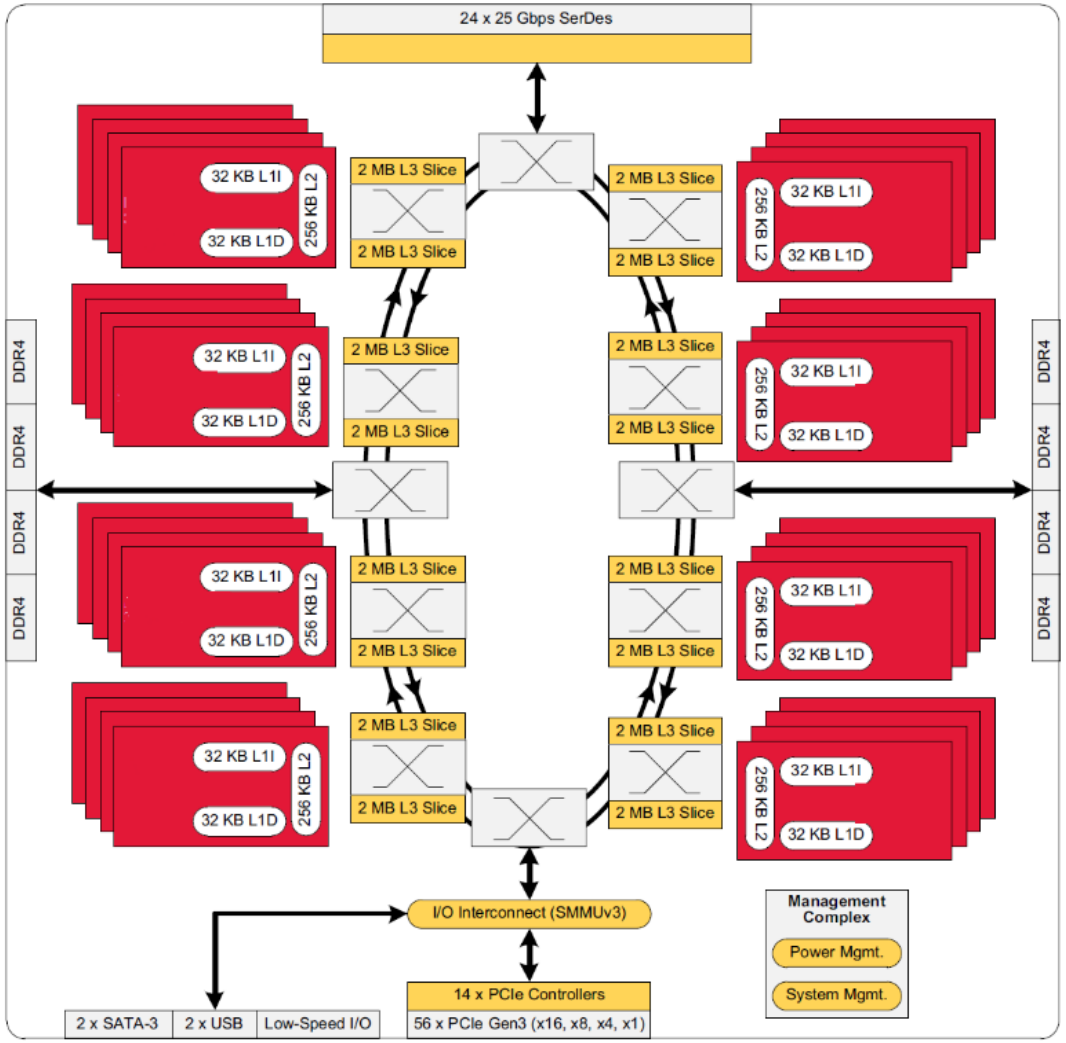
Issues with multicast setup for HCOLL/SHArP in M-OFED.

In the case of both Astra and CatalystUK, I would make the case that issues such as these are exactly what the systems were intended to uncover and drive to resolution.

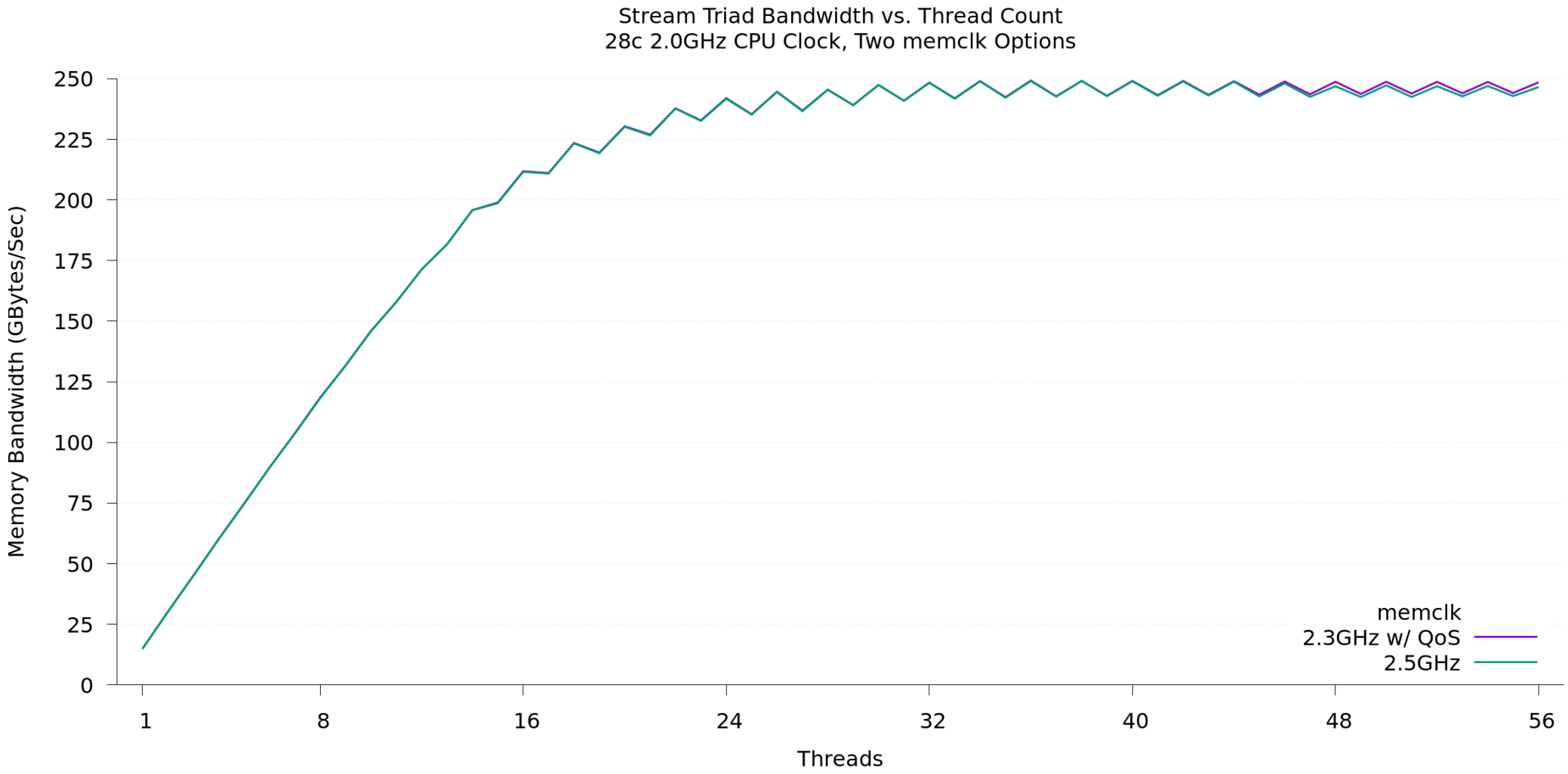
* As specifically requested by Jon Masters from Redhat at the 2017 Going Arm Workshop

Marvell TX2 L3 Throughput

- Single configurable clock used throughout the L3 subsystem (“memclk”)
- Memclk not *directly* related to DIMM speeds
- 2.5GHz used throughout early deployments providing maximum performance with DDR4-2667 memory
- QoS capability designed into silicon to give small messages from GCU to local CPUs priority over cache lines
- Applications routinely decomposed to maximize NUMA-awareness, minimize the inter-socket cache-line traffic, increasing the proportion of NAKs
- Extensive testing demonstrated that memclk could be reduced to 2.3GHz while maintaining maximum memory bandwidth when QoS is enabled
- Power savings and corresponding thermal margin can be redirected to CPU cores, or banked
- 2.3GHz (with QoS enabled) is now the maximum memclk across all TX2 SKUs



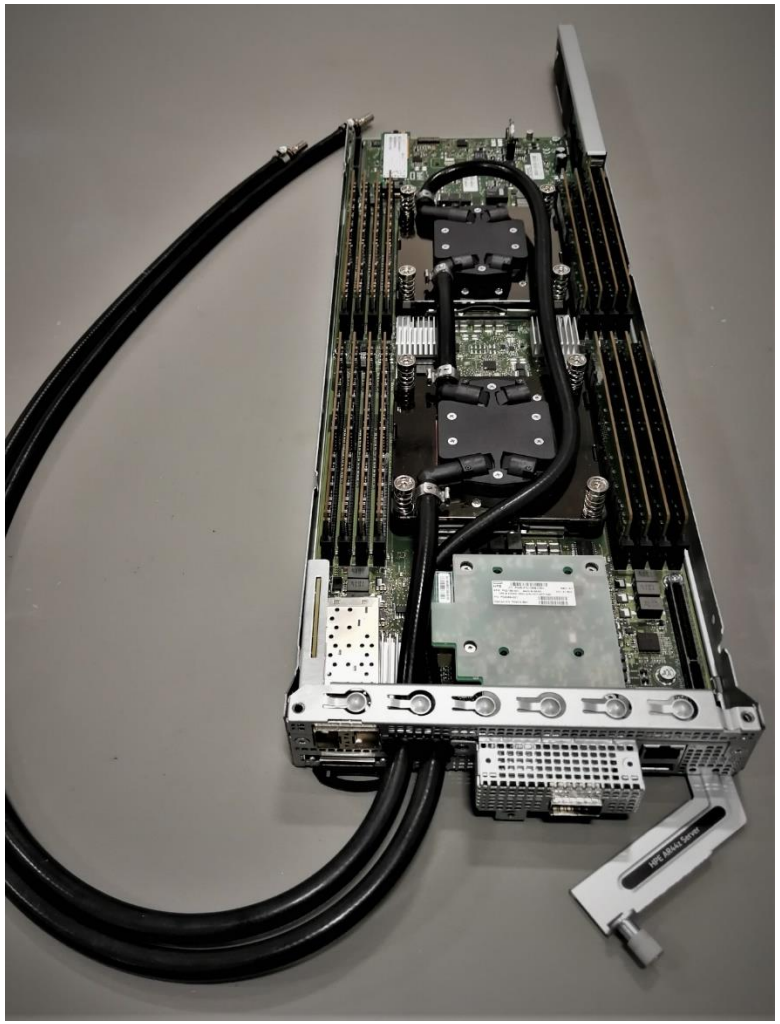
Marvell TX2 L3 Throughput



Liquid-Cooled Apollo 70

- Technology demonstration
- SKUs tested:
  - 28 Core 2.0GHz
  - 32 Core 2.0GHz
  - 32 Core 2.2GHz
  - 32 Core 2.5GHz*
- Modifications limited to CoolIT cold plates and associated plumbing
- Unmodified A2K enclosure, firmware or software
- Displayed in HPE, Marvell & CoolIT booths @ ISC

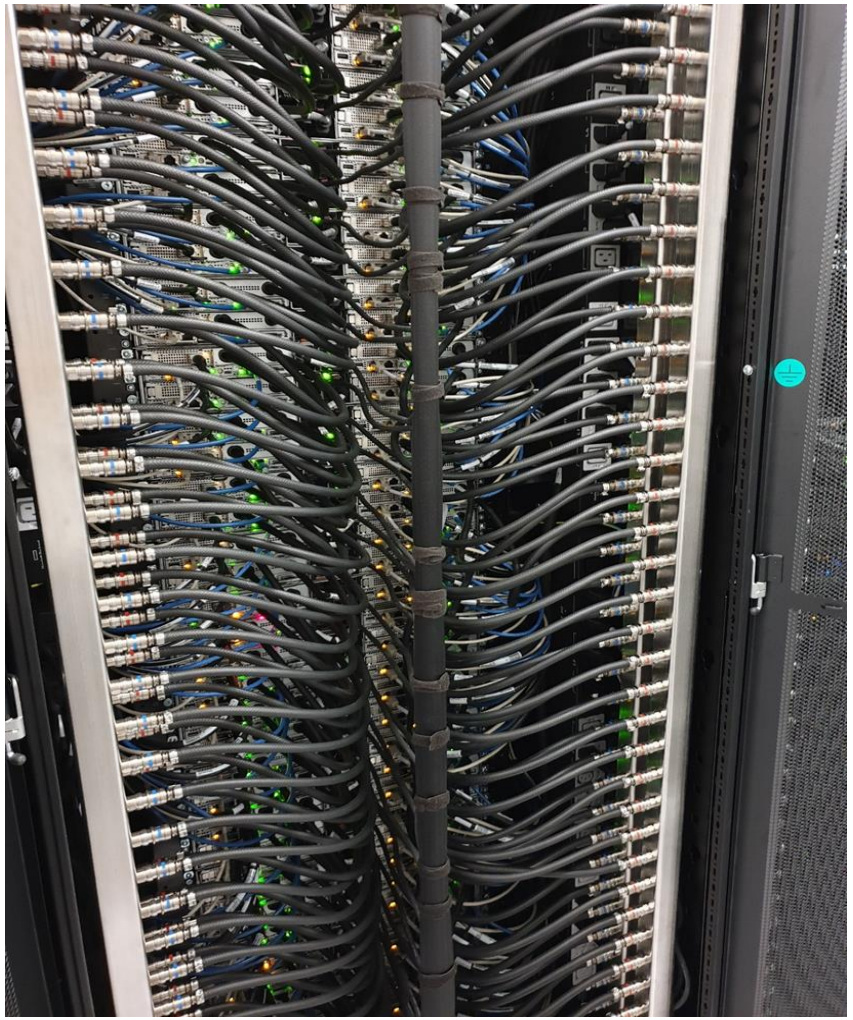
(*) Experimental SKU supplied by Marvell.



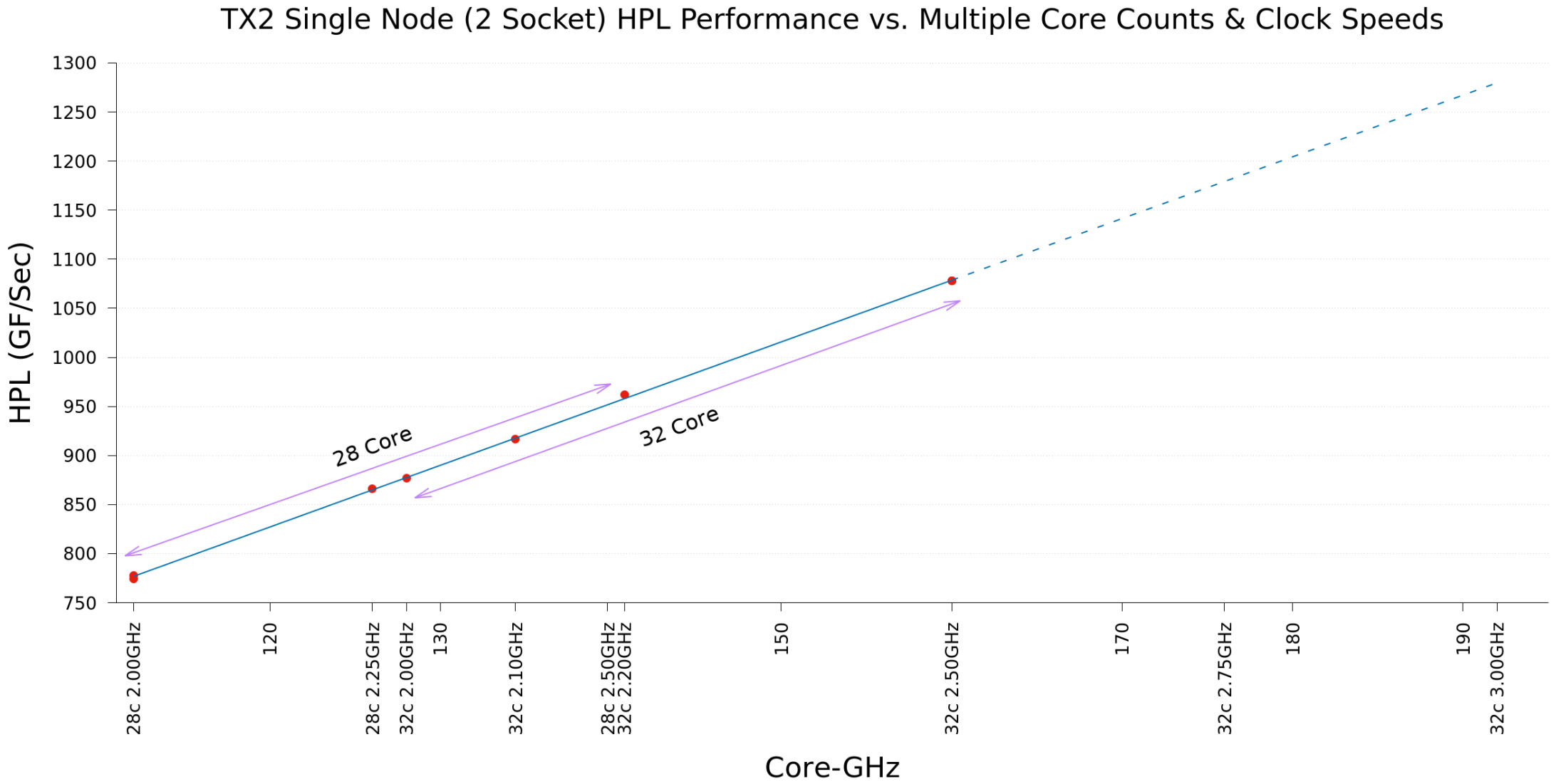
Liquid-Cooled Apollo 70

- Rack level direct liquid cooled Apollo2000 solution‡
- Manifolds installed in rack with non-drip quick disconnects to individual nodes
- Multiple CDU options:
  - Liquid to liquid
  - Liquid to air
  - Single rack
  - Multiple rack

‡ Photograph courtesy of Go Virtual Nordic AB.

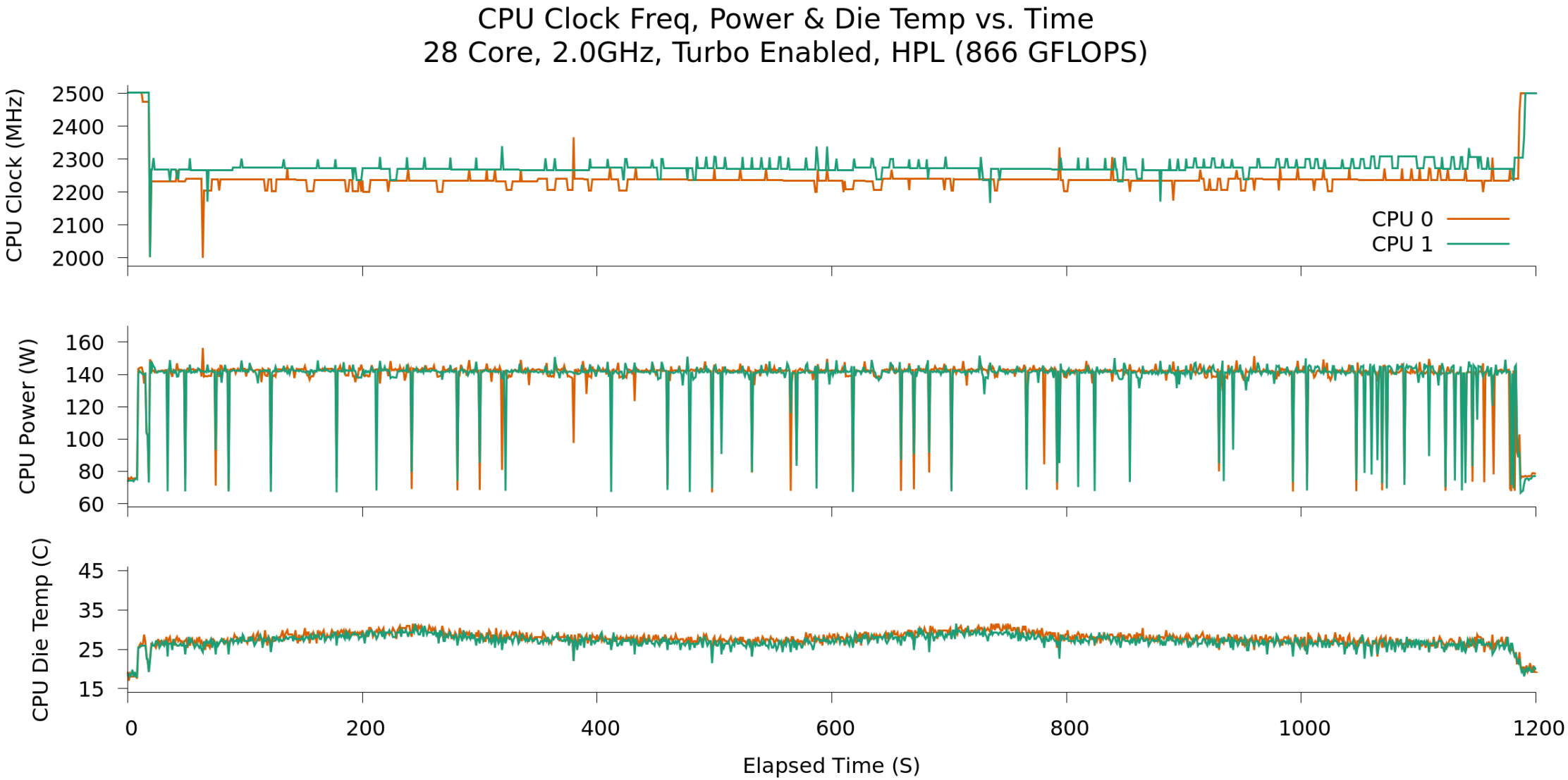


Liquid-Cooled Apollo 70

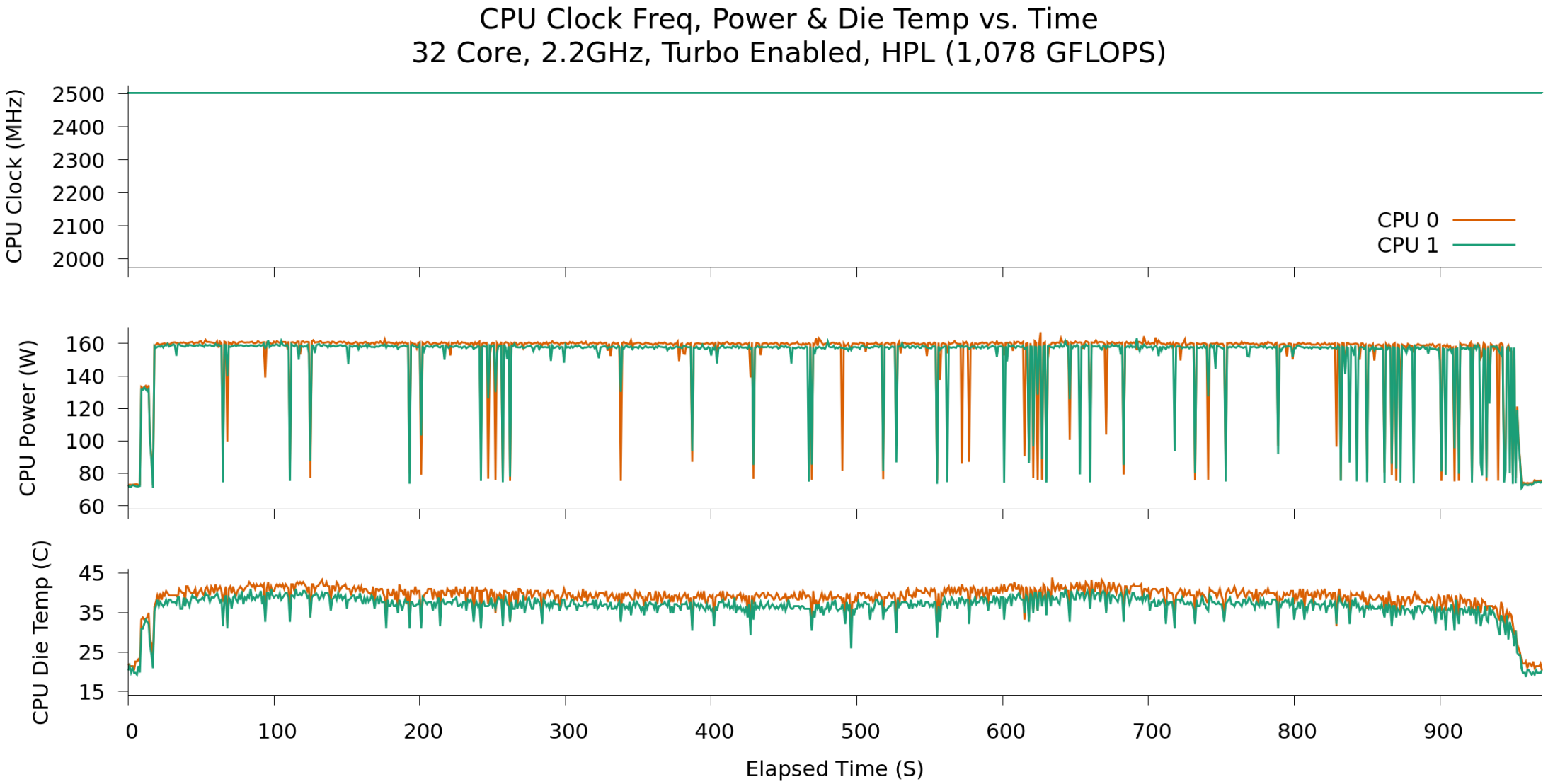


Direct liquid cooling enables full theoretical performance

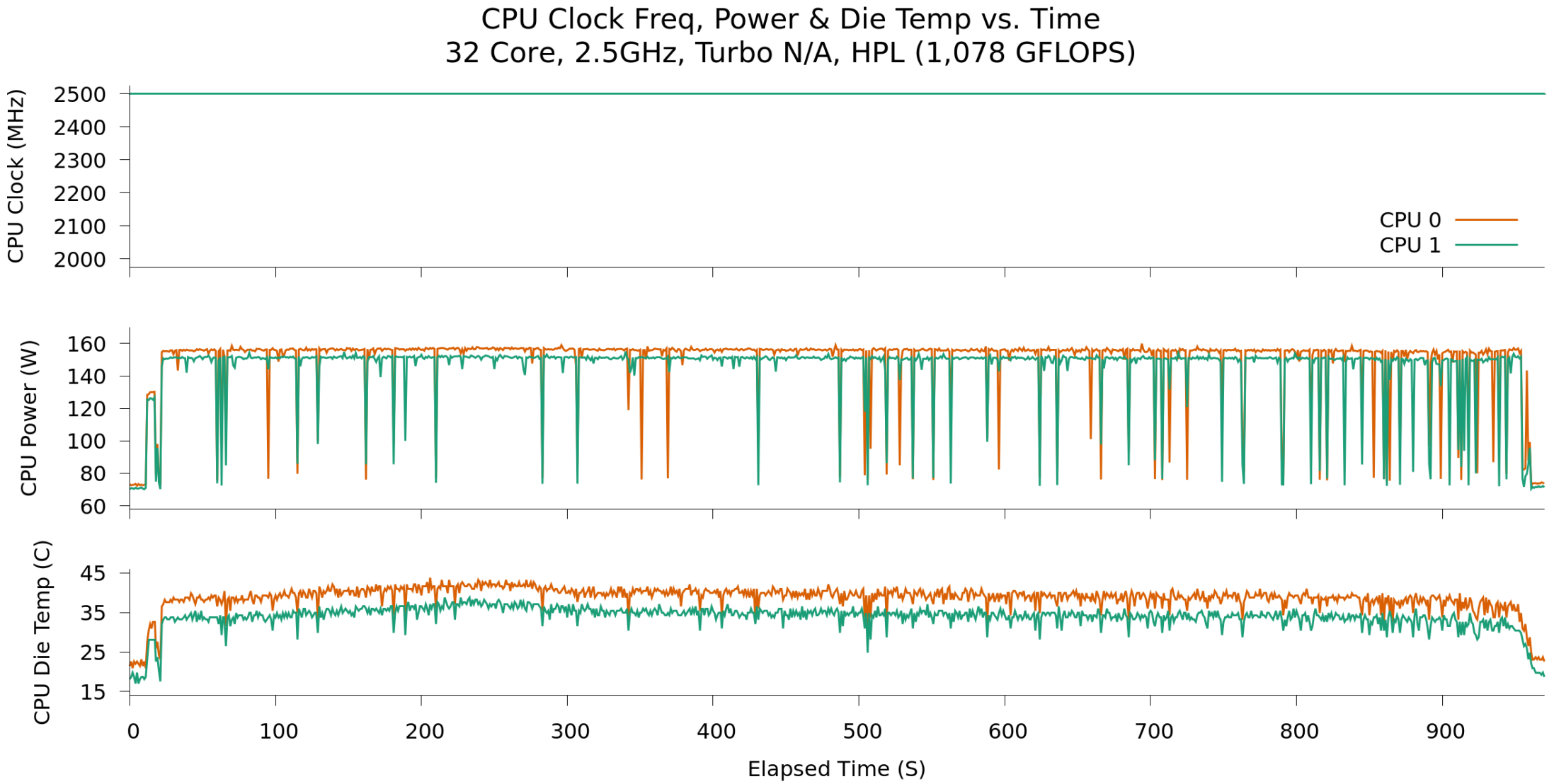
Liquid-Cooled Apollo 70



Liquid-Cooled Apollo 70



Liquid-Cooled Apollo 70



Liquid-Cooled Apollo 70

Conclusions:

- DLC enables continuous operation at max turbo speeds
- Significant headroom available (50ºC thermal, 25-45W electrical power)
- 28c 2.0GHz (150W) SKU performance throttled by power cap
- 32c 2.2GHz SKU performs identically to experimental 2.5GHz SKU due to 100% turbo operation
- Fans @ 40% throughout test, could reduce further with custom firmware

NVIDIA news

NVIDIA

PLATFORMS DEVELOPERS COMMUNITY SHOP DRIVERS SUPPORT ABOUT

NEWSROOM Multimedia Executive Bios Media Contacts In T

News

NVIDIA Brings CUDA to Arm, Enabling New Path to Exascale Supercomputing

Monday, June 17, 2019

Global HPC Leaders Join to Support New Platform

International Supercomputing Conference -- NVIDIA today announced its support for Arm CPUs, providing the high performance computing industry a new path to build extremely energy-efficient, AI-enabled exascale supercomputers.

HPE Apollo70 is GPU ready and available today.

- Up to two GPUs per Apollo70
- PCIe Gen3 x16 links connects each GPU to one CPU
- Two nodes and four GPUs per 2U enclosure