Building a Tier-3 Based on ARMv8 64-bit Server-on-Chip for WLCG

Peter Elmer, Robert Knight - Princeton University
Brian Bockelman - U. Nebraska-Lincoln
David Abdurachmanov, Giulio Eulisse - FNAL

Distributed Computing in HEP - 1990's version

- Heterogeneous computing, workstation era
- Distributed computing as a collection of independent clusters, local job submission
- Many vendors: DEC (VMS, VAX/Alpha), Silicon Graphics, HP, IBM, etc.
- Experiments often supported multiple platforms, as needed, depending on which resources collaborators had

Distributed Computing in HEP - 2000-2015

- High Throughput Computing converged on x86/Linux around ~2000.
- It wasn't because Intel made the most performant processors. Commodity killed off the workstations.
- This convergence was of course a significant simplification which subsequently enabled the WLCG as we know it today: build once, run anywhere.
- (x86-64 was a only minor backwards compatible evolution from x86.)

Evolution of processors

- The commodity market was driven by price/performance, but of course the dirty secret was that power use and density were scaling with performance.

- General purpose processors, superscalar, pipelined, backwards compatible. No widespread interest in specialized functionalities like vectorization.

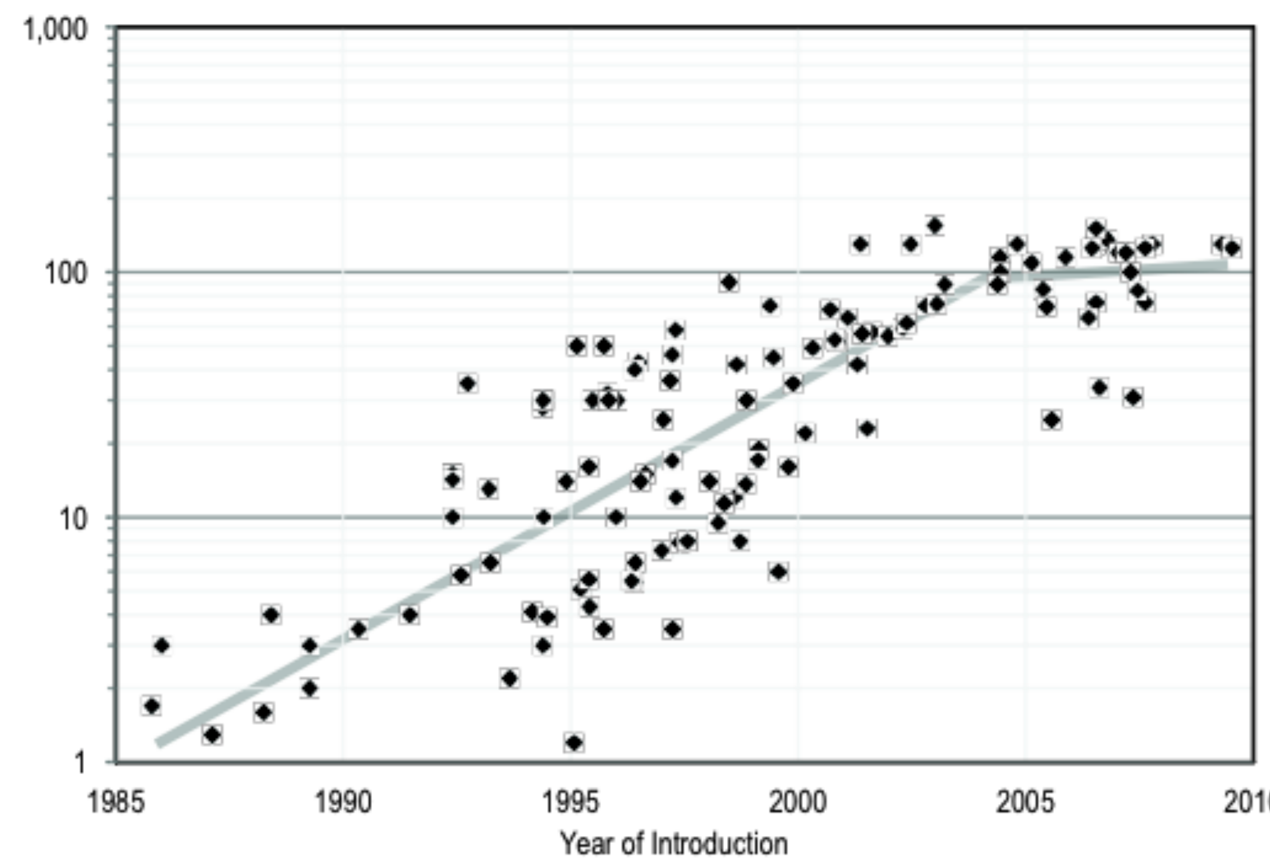


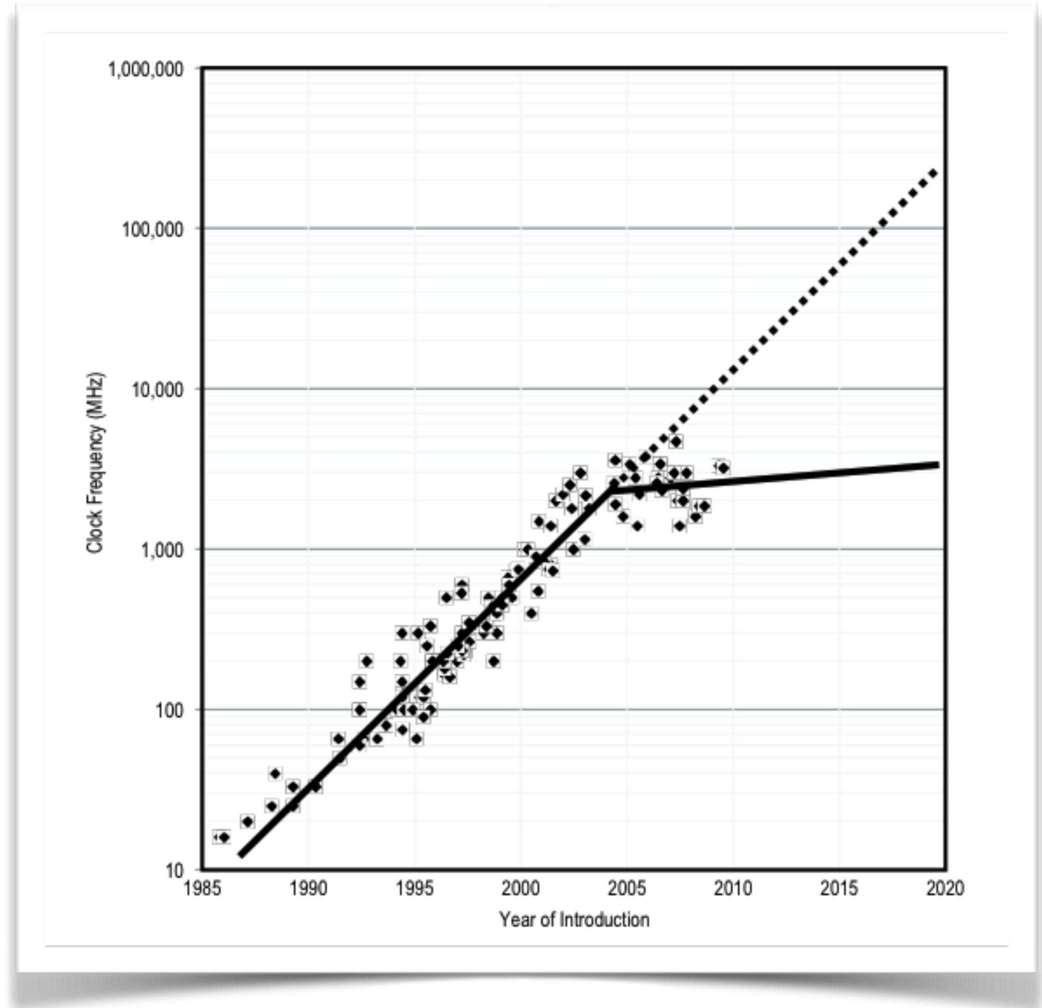
FIGURE A.4 Microprocessor power dissipation (watts) over time (1985-2010).

From: "The Future of Computing Performance: Game Over or Next Level?"

The Future of Moore's Law

- Even multi-core, implemented with large "aggressive" cores is just a stop-gap. The power limitations remain. The focus is shifting to performance/watt, not just performance/price.

- **Overall performance/$$$ growth dropped from 40+%/year before 2005 to 20-25%/year in more recent years**



From: "The Future of Computing Performance: Game Over or Next Level?"

Evolution of processors - Heterogeneous future?

- Given the addition of an overall power limitation, specialized processors and mixes become more interesting: lightweight general purpose cores, vector units, GPUs, mixes like Intel MIC, etc.

- Applications need to adapt and different hardware choices might make more sense for different applications.

- Brute force method of buying the latest/highest performance/$$$ bulk commodity processors is harder.

ARM processors

- In addition to the processor technology evolution, the other large change in the past 10 years is the emergence of mobile computing and cloud computing (as the "server side of mobile")

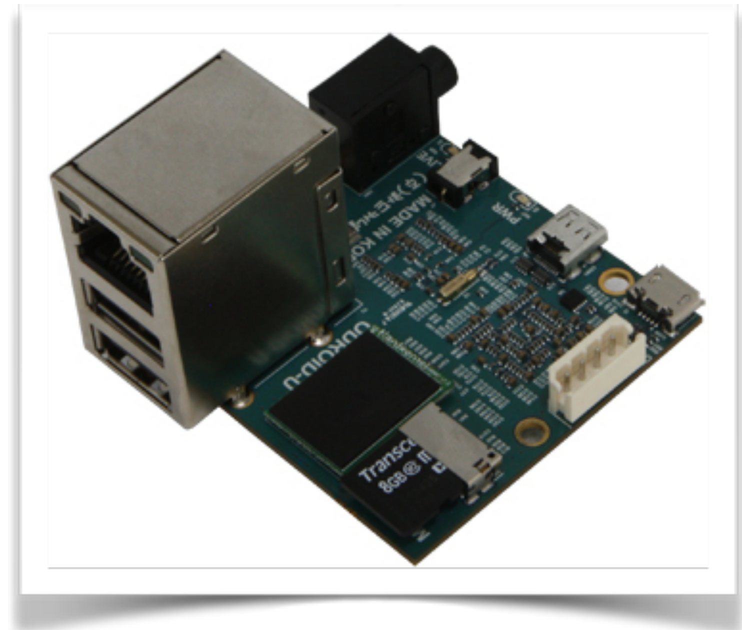
- The mobile/embedded market has been dominated by ARM processors rather than Intel, arguably because of a focus on low power, but also because of a different Intellectual Property model (licensed, rather than produced by ARM) allowing precisely the heterogeneous specialization described earlier.

ARM Processors

- The open question is whether the characteristics that enabled ARM in the mobile market will enable it to compete in the cloud/data center market, dominated by a handful of very big players.
- Another round of commoditization as 15-20 years ago?
- In addition, we see that ARM playing a role in the strategies of other players (AMD, NVIDIA).
- For a couple of years, we have been exploring how ARM could be integrated into our computing systems.

Step 1 - Software port - ODROID U2 (2013)

- Initial software tests were done with a small 32bit/ ARMv7 development board
- Basically a Samsung cell phone chip on a board
- Exynos4412 Prime CPU
- 1.7GHz Cortex-A9 quad core
- 2GB L-DDR memory (total)



- Demonstrated that we can still do a functional port of the whole software stack to non-x86
- Also demonstrated that other performance/power operating points are obtainable

Step 2 - Mustang Board - XGene1 (2014)

- Demonstrated software on server-grade ARMv8 (64bit) development board
- Applied Micro XGene-1 processor
- High performance/power

See talk "Future Computing Platforms ..." by Giulio Eulisse for updated perf/power numbers for different platforms

Heterogeneous Tier-3 demonstrator on OSG

- What is necessary for ARM-based production worker nodes to be a credible alternative to x86-based nodes for use in real sites (given the availability of application level software like CMSSW)?

- The next step is demonstrating that such nodes can be added as a "drop-in" replacement for x86 nodes in grid sites, perhaps even mixed heterogeneously.

- To that end, we have proceeded at Princeton to demonstrate integration of the Mustang board into OSG-style grid access.

Current configuration

- **Worker node:** dagr.princeton.edu (ARMv8 Mustang). Installed - condor, osg-wn-client (globus tools, transfer tools). Runs - condor_startd
- **Batch system head node:** dagr.princeton.edu (also ARMv8 Mustang; typically a separate host). Installed - condor. Runs - condor_collector, condor_negotiator
- **CE:** byggvir.princeton.edu (x86-64). Installed - condor, osg-ce. Runs - condor-ce, condor_schedd
- **CMS specific:** CVMFS (ARMv8 Mustang, client), CVMFS/Frontier squid proxy (x86-64, shared with x86-64 nodes), remote data access via xrootd redirector (Nebraska) and data federation

Demonstrated

- Job submission and execution on ARMv8/64bit worker node
- Previously demonstrated: execution of CMS software on ARMv8, using CVMFS, xrootd, etc.

Issues/Lessons learned

- Private builds of HTCondor, CVMFS and dependencies were done for AArch64 (ARMv8)

- Mustang board runs Fedora19. All issues encountered were related to Fedora19 and are believed to be fixed in Fedora21.

- However an official OSG build will be required for AArch64 on Fedora21 (and perhaps x86-64 on Fedora21).

- HTCondor assumes same arch on submit node and execution node. Manual specification of AArch64 was done. Some testing scripts brought in /bin/sh from submit node, with subsequent exec format errors.

Step 3 - production hardware - HP Moonshot

- High density production hardware based on APM XGene-1 SoC
- Princeton has Moonshot chassis and six m400 server cartridges (each 8 cores, 64GB)
- Ships with Ubuntu-based OS. Currently sorting put OS and firmware issues with Redhat/HP/APM for Fedora/RHEL deployment.
- Technical issue requiring additional network switch in machine location is also being resolved.

Next Steps

- Work towards OSG/HTCondor production builds
- Register the Mustang host with OSG
- Finishing deployment of HP Moonshot system and integrate to OSG as with Mustang board
- Register the HP Moonshot host with CMS for Hammercloud-style tests

Summary

- Power constraints and market evolution may drive change in the kinds of processors we use. Application diversity could drive heterogeneity to aid in {performance, power, cost} optimizations.
- We have been exploring alternatives to the current x86 general purpose cores, including ARMv8/64bit.
- We have demonstrated both application software and grid submission to such nodes, and are in the process of building a small demonstrator cluster.