LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

MNM TEAM
MUNICH NETWORK MANAGEMENT TEAM

lrz

HPC and the AppleTV-Cluster

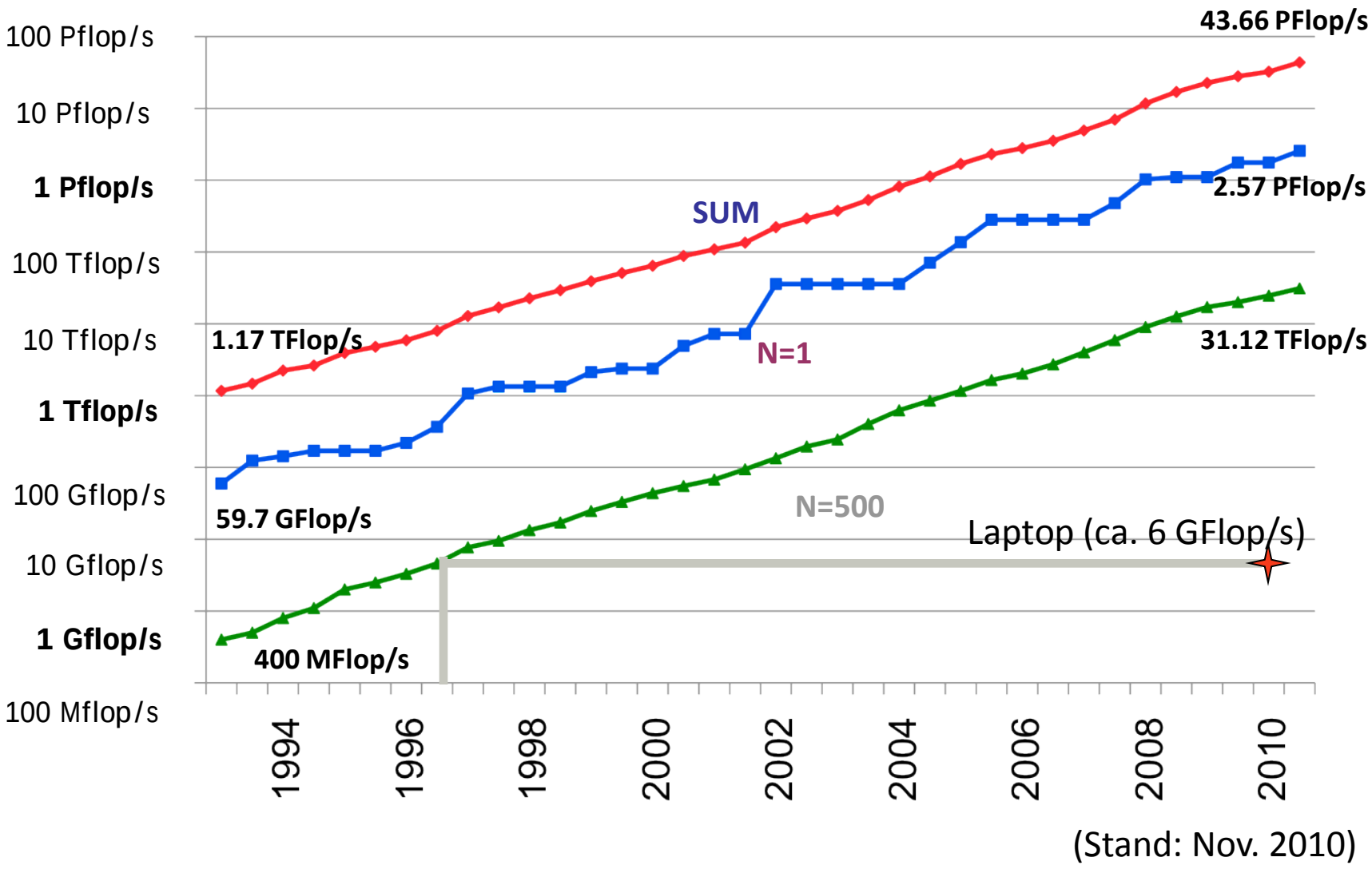
Dieter Kranzlmüller, Karl Fürlinger, Christof Klausecker

Munich Network Management Team
Ludwig-Maximilians-Universität München (LMU) &
Leibniz Supercomputing Centre (LRZ)

Outline

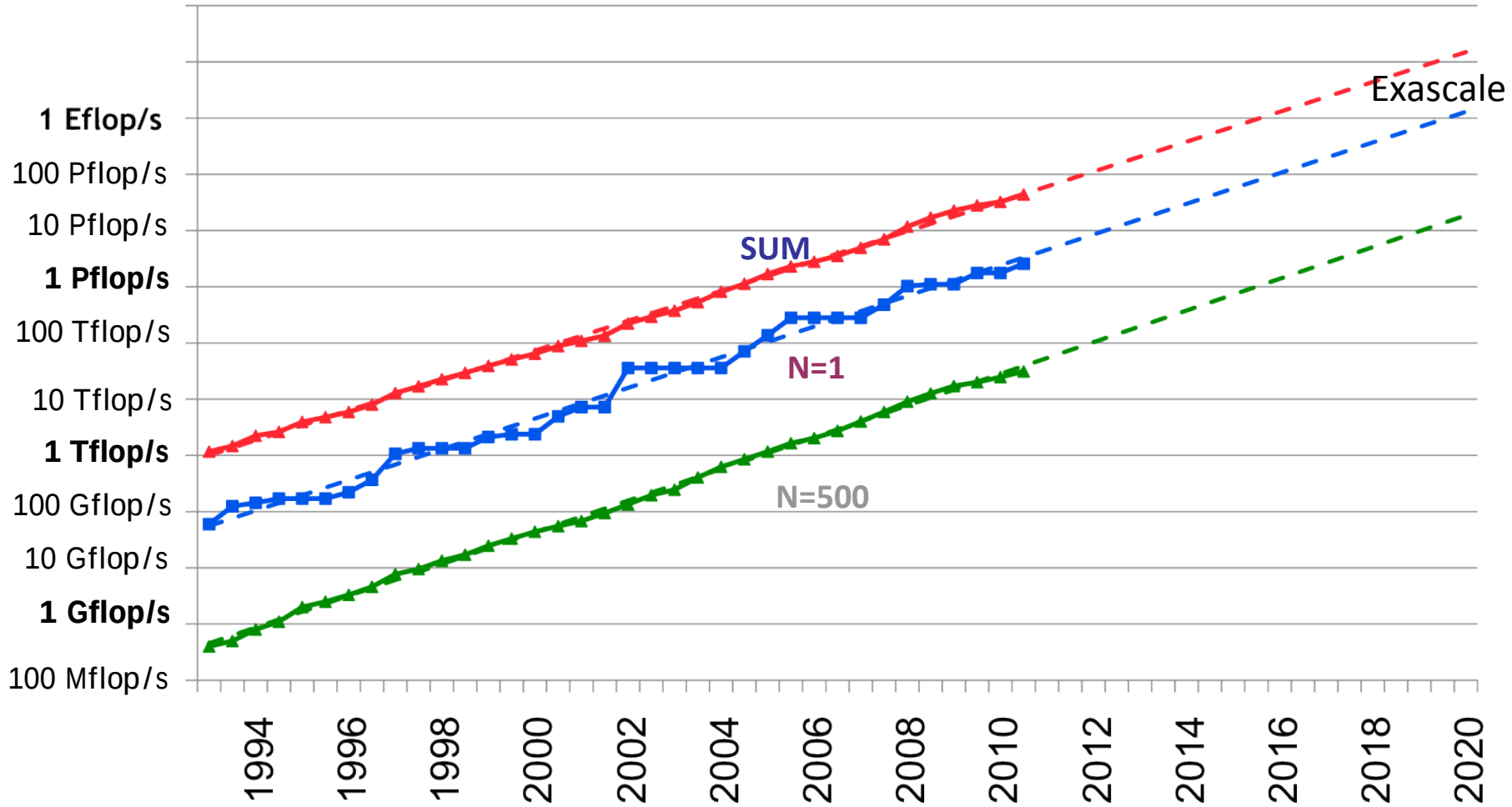
- Motivation
  - Energy efficiency as a primary constraint in HPC
  - The lure of mobile and consumer electronic devices
- The MNM-Team AppleTV-Cluster
  - Hardware and software setup
- Benchmarking
  - Single node results
  - Full cluster performance and power results
- Conclusion

The Current Top500 List



(Stand: Nov. 2010)

Top500 Projection



(Nov. 2010)

Challenges on the Way to Exascale

- Parallelism
  - Extreme scale (billion-way)
  - Limitations of the bulk-synchronous model
  - Programming for heterogeneity
  - Noise and system variablility
- Resiliency, fault tolerance
  - Dropping MTTI
  - Limitations of checkpoint-restart
- Power consumption, energy efficiency
  - 1 MW ~ 1 Mio USD/year (ca 0.11 USD per kWh)
  - Energy efficiency has to improve dramatically
  - Max. 20 MW for an exascale class machine

Energy Efficiency

- **Wikipedia:** "**Efficient energy use**, sometimes simply called **energy efficiency**, is the goal of efforts to reduce the amount of energy required to provide products and services."

- **Energy efficiency in scientific and technical computing**
  - Use the least amount of energy to solve a scientific problem
  - Approximation: MFlops per Watt
  - PUE: Power Usage Effectiveness
- **Green500 list** (www.green500.org)
  - Tracks the power requirement of systems in the Top500 list to solve the linpack benchmark

Green500 List (June 2011)

Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	2097.19	IBM Thomas J. Watson Research Center	NNSA/SC Blue Gene/Q Prototype 2	40.95
2	1684.20	IBM Thomas J. Watson Research Center	NNSA/SC Blue Gene/Q Prototype 1	38.80
3	1375.88	Nagasaki University	DEGIMA Cluster, Intel i5, ATI Radeon GPU, Infiniband QDR	34.24
4	958.35	GSIC Center, Tokyo Institute of Technology	HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows	1243.80
5	891.88	CINECA / SCS - SuperComputing Solution	iDataPlex DX360M3, Xeon 2.4, nVidia GPU, Infiniband	160.00
6	824.56	RIKEN Advanced Institute for Computational Science (AICS)	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect	9898.56
7	773.38	Forschungszentrum Juelich (FZJ)	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus	57.54
8	773.38	Universitaet Regensburg	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus	57.54
9	773.38	Universitaet Wuppertal	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus	57.54
10	718.13	Universitaet Frankfurt	Supermicro Cluster, QC Opteron 2.1 GHz, ATI Radeon GPU, Infiniband	416.78

* Performance data obtained from publicly available sources including TOP500

Source: http://www.green500.org

Energy Efficiency in the Green500 List

Cray XT line of systems

System	MFLOPS/Watt
Cray XT3 (2004)	60
Cray XT4 (2006)	130
Cray XT5 (2007)	150
Cray XT6 (2009)	260
Cray XE6 (2010)	360

Blue Gene line of systems

System	MFLOPS/Watt
IBM Blue Gene/L (2005)	204
IBM Blue Gene/P (2007)	370
IBM Blue Gene/Q*(2011)	2097

*: Prototype; Source: Green500 list, June 2011

- However, Exascale requires at least **50 000** MFlops/Watt (a 20 MW envelope)!
- Dramatic improvements are required to achieve this level of energy efficiency
  - Questionable if evolution of conventional CPUs can achieve this
- Revolutionary approaches:
  - HW/SW codesign
  - Try something different…

Mobile and Consumer Electronic Devices

- Energy efficiency was always a primary concern in the mobile area
- Use cases are getting more computationally demanding
  - HD video encoding/decoding, augmented reality, rich web applications
- Devices are full-featured computers
  - Often with a UNIX-like OSs, 100s of MBs of RAM, GBs of storage
  - Dual-/Quadcore CPUs are appearing

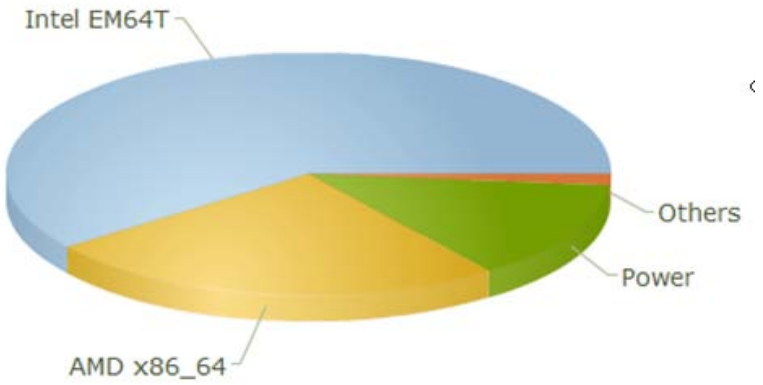
NVIDIA

Whitepaper

The Benefits of Multiple CPU Cores in Mobile Devices

Energy Efficient Parallel Computing on Mobile Devices?

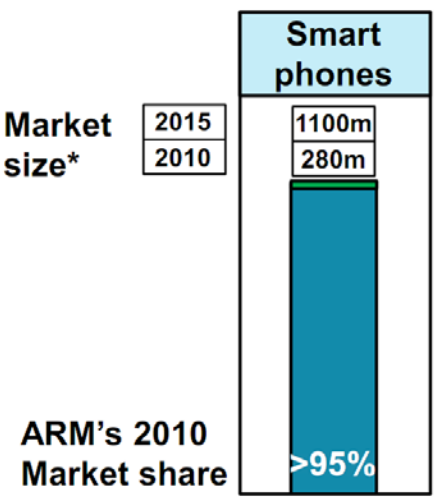
- The HPC ecosystem



Processor Family / Performance
November 2010 (Top 500)

- 90% of systems in the Top500 use Intel/AMD CPUs
  - Supercomputer market is too small to sustain custom CPU designs
  - Supercomputer vendors (except IBM) rely on server/desktop CPUs and GPUs

- The mobile computing ecosystem



Source: ARM Holdings Q1/2011 financial results report

Outline

- Motivation
  - Energy efficiency as a primary constraint in HPC
  - The lure of mobile and consumer electronic devices
- The MNM-Team AppleTV-Cluster
  - Hardware and software setup
- Benchmarking
  - Single node results
  - Full cluster performance and power results
- Conclusion

The AppleTV Cluster

- Platform: 2nd Generation AppleTV (ATV2)

- Small
  - 10x10x2.3 cm, 270g
- Cheap
  - MSRP 100,- USD
- Powerful
  - Same HW as iPad (1st gen.)
  - Apple A4 processor @1GHz
    ➔ ARM Cortex-A8 CPU+PowerVR GPU
  - 256 MB RAM
  - 8 GB NAND Flash
- „Green“
  - 1-3 Watts

The MNM-Team AppleTV Cluster

- We‘ve built a small cluster of ATV2 nodes
  - 4+2 nodes, connected by a 100 MBit Ethernet switch

Software on the ATV cluster

- BSD-based OS: iOS 4.2.1 (Darwin kernel version 10.4.0)
- *Jailbreak* necessary to install custom software
  - Many choices provided by the iOS jailbreak community
  - We‘ve used greenp0ison
- System with running ssh server after jailbreak with root access
  - apt-get install,…
  - See a small howto guide on our webpage (www.appletvcluster.com)
- Editors, gcc toolchain, …
  - Gcc 4.2.1
- MPI
  - MPICH 2 from Argonne National Lab
  - hydra process manager
  - TCP transport

Outline

- Motivation
  - Energy efficiency as a primary constraint in HPC
  - The lure of mobile and consumer electronic devices
- The MNM-Team AppleTV-Cluster
  - Hardware and software setup
- Benchmarking
  - Single node results
  - Full cluster performance and power results
- Conclusion

Benchmarking

- Single node benchmarks
  - Memory system performance
  - CPU performance
- Whole cluster benchmarks
  - MPI microbenchmarks
  - Linpack
- …this is ongoing work
  - We‘re reporting initial results here
  - Work left to be done and extended

Single Node Performance

- Two metrics:
  - Direct comparison with a **BeagleBoard**
  - (Open source ARM-based development platform)
  - Orders of magnitude comparison with conventional server CPUs

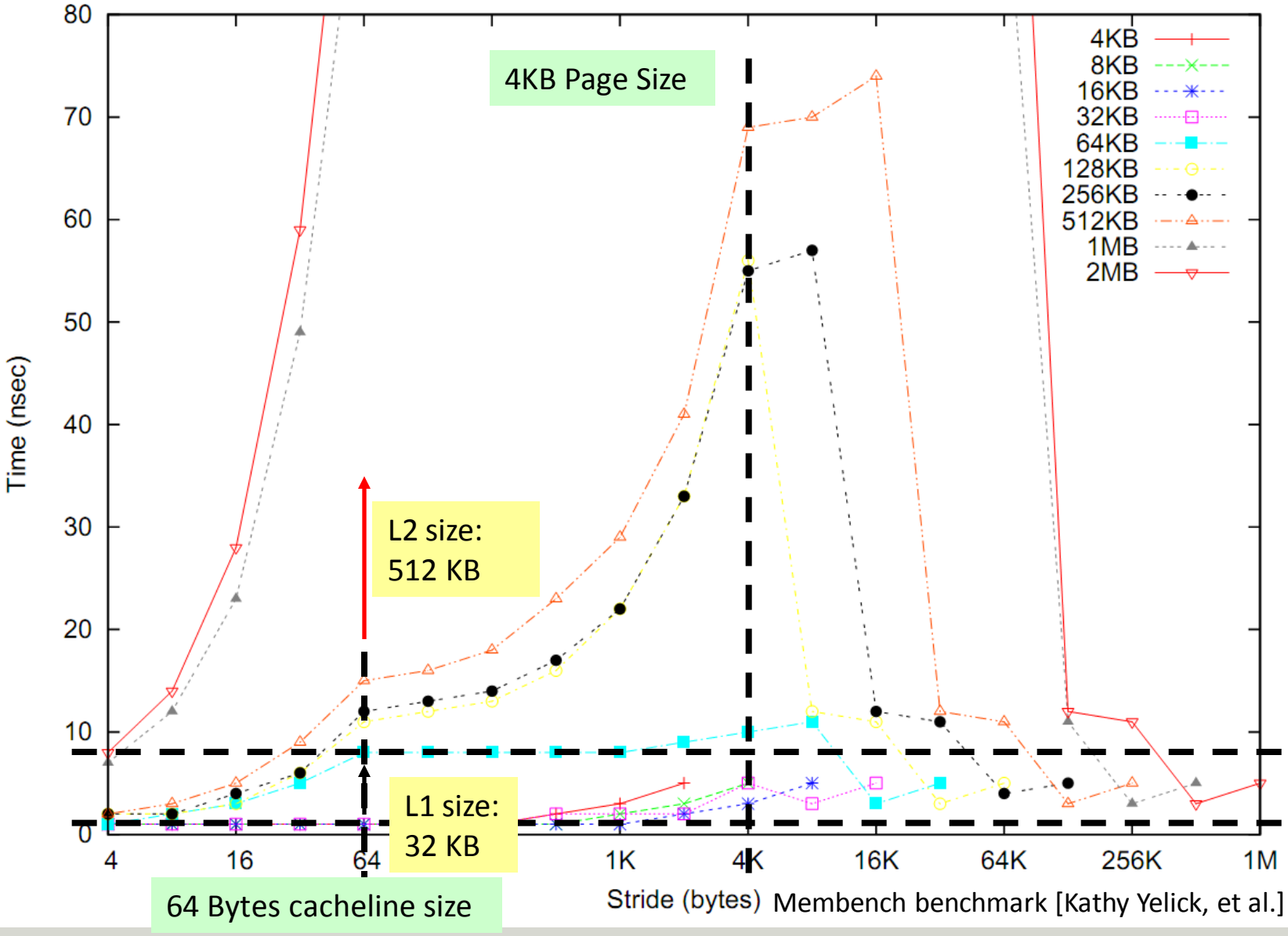
- BeagleBoard-xM
  - TI DM3730 Processor - 1 GHz Cortex-A8 [our BeagleBoard-xM was running at 800 MHz]
  - 512 MB LPDDR RAM memory

Node Parameters

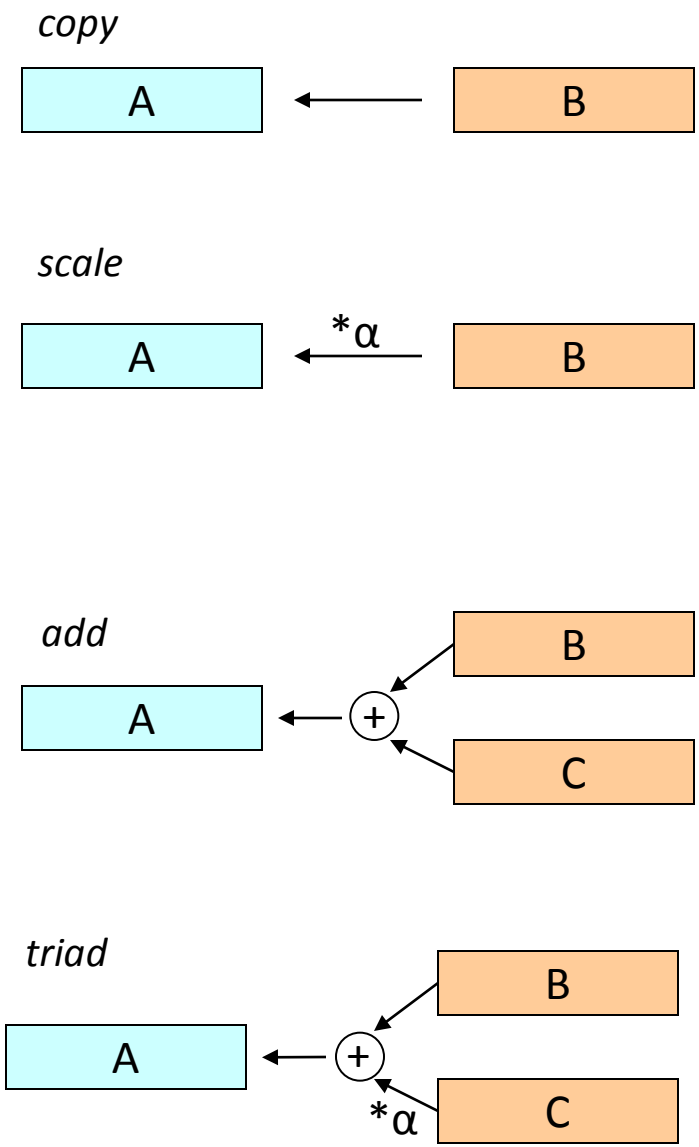
Parameter	Value	Source
Core frequency	1 GHz	Experiments
L1I cache size	32 KB	hw.l1icachesize
L1D cache size	32 KB	hw.l1dcachesize
L2 cache size	512 KB	hw.l2cachesize
L1 latency	1 cycle	Experiments
L2 latency	8 cycles	Experiments
Cache line size	64 B	hw.cachelinesize
Bus frequency	100 MHz	hw.busfrequency
Memory size	247 MB	hw.memsize
Page size	4 KB	hw.pagesize

- Cortex-A8
  - In-order, super-scalar processor (dual issue)
  - Compare to Intel Pentium III-S with approximately similar characteristics (ca. 2001)

Memory System Parameters and Performance



Stream (Memory Bandwidth)



Operation	**BeagleBoard** xM (800 MHz) [MB/s]	**ATV2** (1 GHz) [MB/s]
copy	481.1	749.8 (+55%)
scale	492.9	690.0 (+40%)
add	485.5	874.7 (+80%)
triad	430.0	696.1 (+60%)

- ATV2
  - 200 MHz, 64 bit memory bus?
- BeagleBoard
  - 166 MHz, 32 bit
- Server: Core i7, 800 MHz DDR2
  - ~ more than 10 times higher bandwidth

Coremark (Integer Performance)

- EEMBC benchmark to test processor core performance
  - Similar in purpose to Dhrystone
  - Integer, control, memory operations
  - Small binary footprint

Device	Absolte Coremark Score	Coremark / MHz
BeagleBoard xM (800 MHz)	1928	2.41
ATV2 (1 GHz)	2316	2.32

- Both platforms have a Cortex-A8 CPU
  - Essentially the same performance per MHz

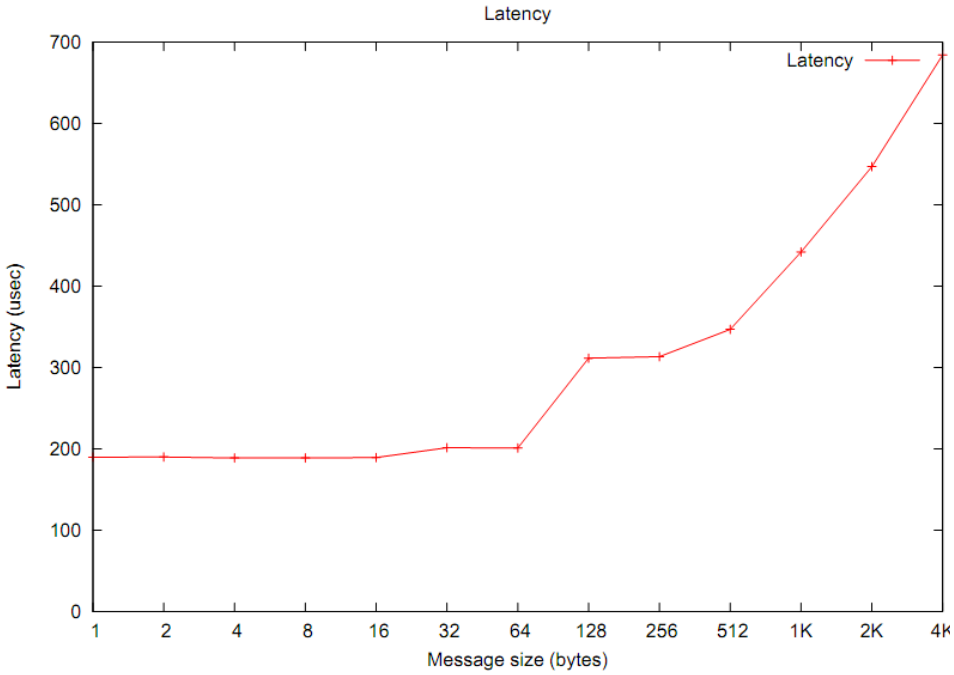
- Comparing Coremark scores from Desktop/Server CPUs:
  - High absolute performance but small difference when normalized to Perf/MHz/Thread

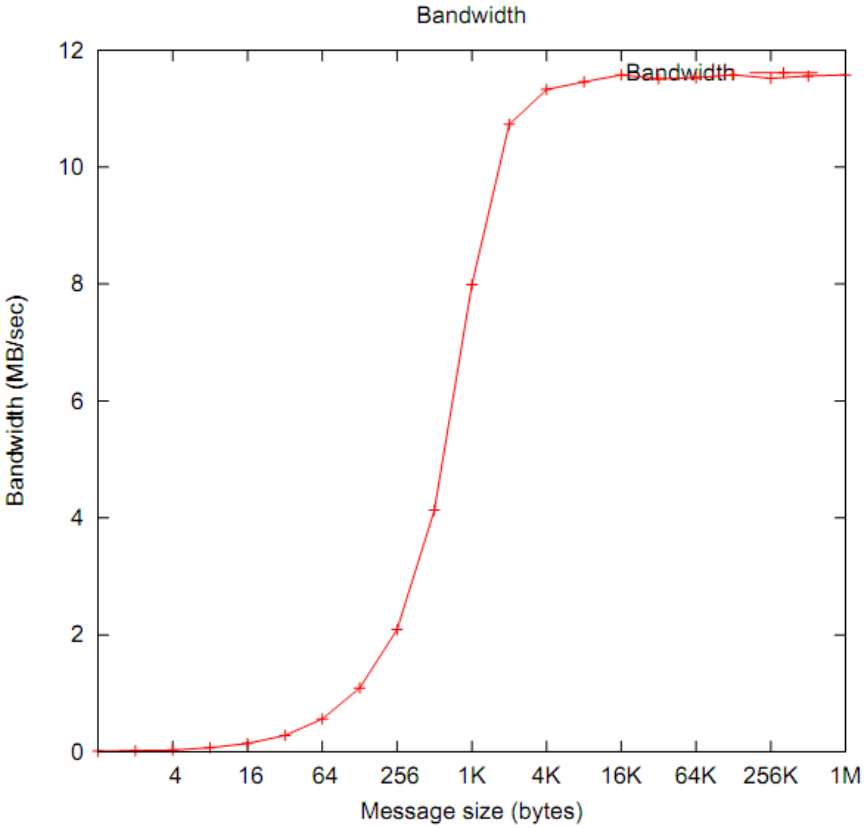
Outline

- Motivation
  - Energy efficiency as a primary constraint in HPC
  - The lure of mobile and consumer electronic devices
- The MNM-Team AppleTV-Cluster
  - Hardware and software setup
- Benchmarking
  - Single node results
  - Full cluster performance and power results
- Conclusion

Cluster Performance

- MPI Ping-Pong Latency and Bandwidth measurements
  - 100 Mbit Ethernet and TCP as the transport mechanism
  - Compare to high perf. Interconnects (IB QDR / 10GigE Ethernet): **2-6** usec Latency sec 1 GB/sec Bandwidth
  - → Two orders of magnitude slower





Floating Point Performance

- Linpack benchmark
  - Solves *Ax=b* (dense matrices), using LU factorization
  - R_{max} value is the basis for the Top500 ranking
  - High comp. intensitiy, systems achieve 60-90% of peak

- ATV2 cluster result
  - 4 nodes achieve R_{max} of **160.4** MFlops/Watt
  - Power consumption: about 10 Watts (all 4 nodes)
  - → Energy efficiency of 16 MFlops/Watt
  - Compare Green500 list:
    - #1: 2097 MFlops/Watt
    - #500: 21 MFlops/Watt

Floating Point Performance - Analysis

- Cortex-A8 really *not* optimized for DP floating point performance
  - SIMD unit (NEON) only for SP FP numbers
  - DP numbers take VFP execution path
  - VFP execution path isn‘t even pipelined
  - DP FP operation latency is 9-17 cycles
  - → Peak DP performance is probably 60-70 MFlops
- What‘s ahead:
  - Cortex-A9 (dual-core): iPad2, iPhone 4S, Samsung Galaxy-Tab, …
    - Much improved floating point performance
  - Cortex-A15 (2012, 4+ cores, out-of-order, superscalar)
    - Expected to compete with desktop CPUs

Cortex-A9: Adoption in the HPC Area

Dongarra Runs LINPACK on the iPad2

05.10.2011

Like 7 John Markoff of the NY Times writes that Jack Dongarra has run LINPACK on an iPad2. In fact, he obtained performance results that would rival for the CRAY-2 supercomputer, which was the world's fastest machine in 1985.

"Dr. Dongarra's researchers also discovered that the new iPad2 is about 10 times as fast as its predecessor, the original iPad. That is likely because of some design changes in the microprocessor used in the new version of the Apple tablet. To date, the researchers have run the test on only one of the iPad microprocessor's two processing cores. When they finish their project, though, Dr. Dongarra estimates that the iPad 2 will have a Linpack benchmark of between 1.5 and 1.65 gigaflops (billions of floating-point, or mathematical, operations per second). That would have insured that the iPad 2 could have stayed on the list of the world's fastest supercomputers through 1994.

(NY Times, via insidehpc.org)

SEVENTH FRAMEWORK PROGRAMME
THEME ICT-2009.9.13
Exa-scale computing, software and simulation

SEVENTH FRAMEWORK PROGRAMME

Proposal acronym:	Mont-Blanc[1]
Proposal full title:	Mont-Blanc, European scalable and power efficient HPC platform based on low-power embedded technology

ARM Cortex-A9 performance in HPC applications

M. Boyd [1], C. Della Silva [2], K. Keville [3]
[1] Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA
mboyd@mit.edu
[2] Department of Mechanical Engineering, MIT, Cambridge, MA
clarkds@mit.edu
[3] Institute for Soldier Nanotechnologies, MIT, Cambridge, MA
kkeville@mit.edu

SoC/Dev Board	Ti Pandaboard EA3	C
CPU / GPU	Cortex A-9/ SGX540	A
L1 Cache Size	32 KB	1
L2 Cache Size	512 KB	2
CPU Speed	1GHz	6
RAM Memory	1GB DDR2	2
L1I Cache Size	32 KB	1
L1D Cache Size	32 KB	1
L2 Cache Size	512 KB	2
LINPACK SP	3.0 GFLOPS	2
LINPACK DP	1.2 GFLOPS	1
Whetstone single core	1376 MIPS	1
Dhrystone single core	1824 DMIPS	9
Coremark (iter / sec)	5637 (dual core) 2816 (single core)	1

Outline

- Motivation
  - Energy efficiency as a primary constraint in HPC
  - The lure of mobile and consumer electronic devices
- The MNM-Team AppleTV-Cluster
  - Hardware and software setup
- Benchmarking
  - Single node results
  - Full cluster performance and power results
- Conclusion

Conclusion

- Mobile devices are full-featured general purpose computers today
  - Running UNIX operating systems
  - GHz CPUs, hundreds of MB or RAM, GBs of storage
  - Dual and quad-core CPU designs are appearing
  - Powerful integrated GPUs
  - Energy efficiency always a primary design consideration
  - iPad2 more powerful than #500 system in 1993

- On the ATV2 cluster (Cortex-A8 based)
  - Performance for FP intensive HPC applications is not competitive ***today***
  - Integer performance is much better
  - No high performance support for DP floating point arithmetic
  - No SIMD for double precision in NEON
  - No ECC protection for memory or caches
  - No high performance interconnects

...but

- Mobile industry is on a steep technology trajectory
  - Programmable integrated graphics cards (OpenCL)
  - Dual/Quadcore CPUs
  - Power-hungry use cases
  - Big momentum behind ARM: NVIDIA, Microsoft are ARM licensees
  - ARM for servers: Calxeda, and others
- Our plans:
  - More benchmarks
  - Integer-intensive and datacenter workloads
  - PandaBoard (Cortex-A9)

http://www.appletvcluster.com

- HOWTO guide
- Whitepaper (with current results))
- Dieter.Kranzlmueller@nm.ifi.lmu.de
- Karl.Fuerlinger@nm.ifi.lmu.de
- Christof.Klausecker@nm.ifi.lmu.de