Millicomputing – The Coolest CPUs and the FLASHIEST Disks

The Future of Computing is Open Hardware by the milliWatt!

http://www.millicomputing.com
acockcroft@netflix.com

October 9, 2007 **HPTS**

Although the author is employed by Netflix Inc. these are the personal opinions of the author and no endorsement by Netflix Inc. is implied.

Agenda

- Enterprise Computing Market Patterns
- Millicomputers
- Milliclusters
- Flash Storage
- Packaging
- Application Implications
- Management Implications
- Next Steps

NETFLIX™

Enterprise Computing

- A repeating pattern
  - Mainframes replaced by Minicomputers
  - Minicomputers replaced by RISC servers
  - RISC servers replaced by PC servers
- The same objections every time…
  - “It’s a toy, not enterprise-ready”
  - “It can’t do big I/O”, “It doesn’t have big memory”
  - “Its more efficient to manage fewer bigger machines”
- What replaces the PC server?
- How do we build “green” datacenters?

NETFLIX™

How to Get $Billion Revenue

- Commodities and $Billion server products
  - 1970: Sell 100 Mainframes @ $10M each
  - 1980: Sell 1,000 Minicomputers @ $1M
  - 1990: Sell 10,000 RISC servers @ $100K
  - 2000: Sell 100,000 PC servers @ $10K
  - 2007: Racked Blades: 50,000 Chassis @ $20K
    - 10 Blades per chassis @ $2K each
  - 2007: Mobile: 10,000,000 Millicomputers @ $100
    - OK for consumer cellphone market, not for enterprise
    - Not economical to sell individually

NETFLIX™

Millicomputers

- Battery powered portable games, phones, cameras etc. have increasingly powerful CPUs inside
- Very low power use, very low cost 32bit RISC with Linux support
- Tiny and reliable system on a chip

- Millicomputer definition
  - A computer that uses less than one Watt

NETFLIX™

The Millicomputing Questions

- Do these CPUs have enough capacity to be useful for general purpose enterprise computing tasks?
- What is the growth trend for millicomputers?
- What is the price/performance, Watts/performance, rack density?
- How hard is it to make a millicomputer?
- How can vendors package tiny cheap machines into products?

NETFLIX

What is Open Hardware? No vendor!

- Which vendor “owns” Linux? Lets make hardware open as well…
  - Fully customizable to fit whatever needs you have!
- Open specifications for components
  - No NDA required for access to the full device specification
  - Allows open source Linux device drivers to be released
- Openly published schematics
  - Free access to circuit designs for open components
  - Supports incremental innovation and design improvements
- Openly published Printed Circuit Board (PCB) layouts
  - Free access to a range of PCB designs and cost tradeoffs
  - Supports incremental innovation and design improvements
  - Easy re-purposing, custom PCB shapes, added devices
- Openly published mechanical packaging
  - 3D CAD files for components and assemblies openly published
  - Free access to a range of mutable layouts and case designs

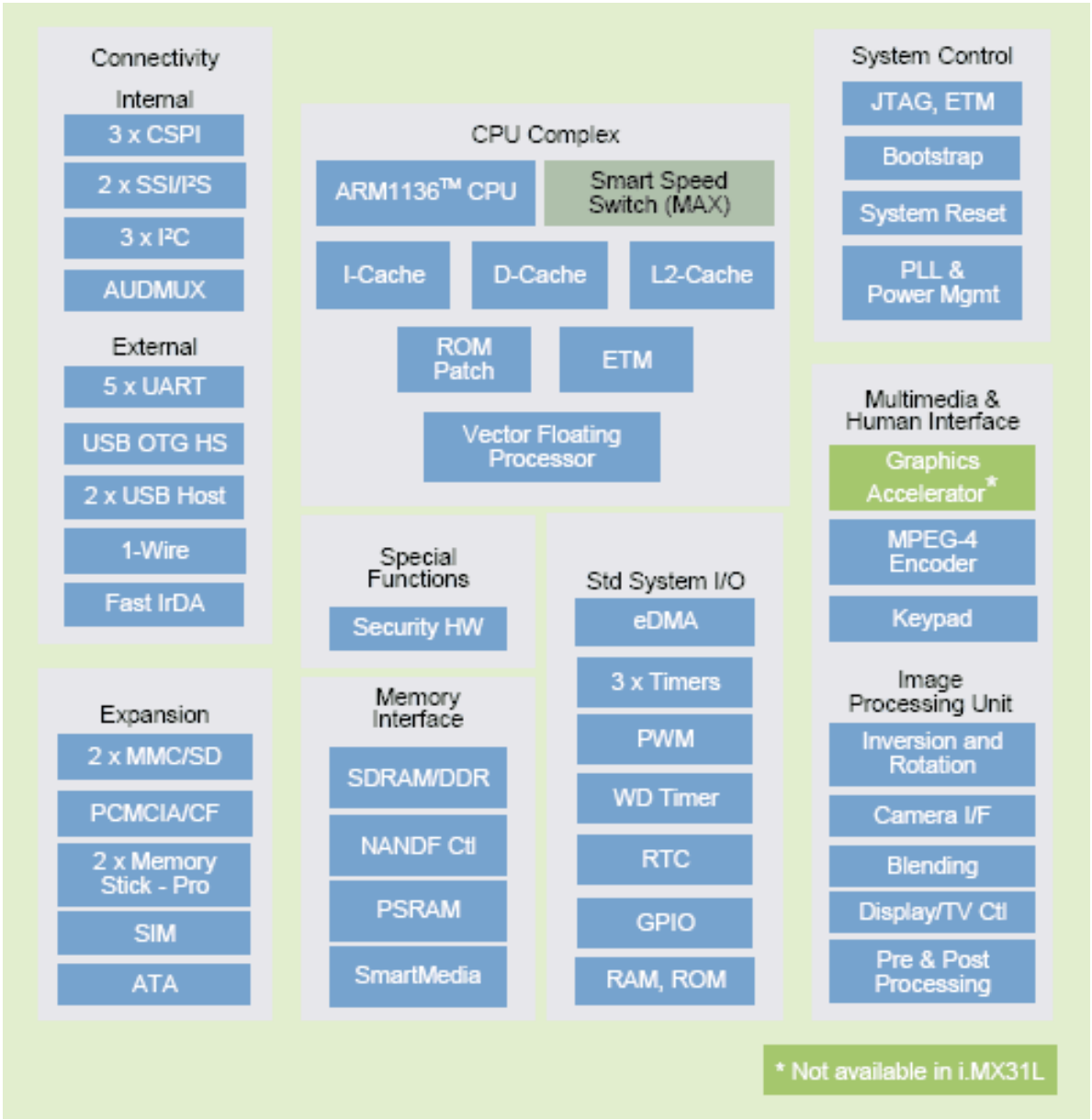
NETFLIX™

Millicomputer's - The Coolest CPUs

- ARM processor architecture – 32bit RISC
  - Power consumption when idle - a few milliwatts
  - Maximum power consumption - 250mW to 900mW
- Vendors – Marvell, Freescale, Samsung
  - Intel XScale business was sold to Marvell
  - Very common use in mobile devices
  - e.g. iPhone, Treo, Zune, iPod, most mobile phones
  - Annual worldwide sales of billions of units
- High End Millicomputer System-on-a-Chip's for ~$20
  - Freescale i.MX31 – 532MHz
    - Includes FPU, Multimedia and 3D acceleration for peak 250mW
  - Marvell PXA270 – 624MHz
  - Marvell PXA320 – 806MHz

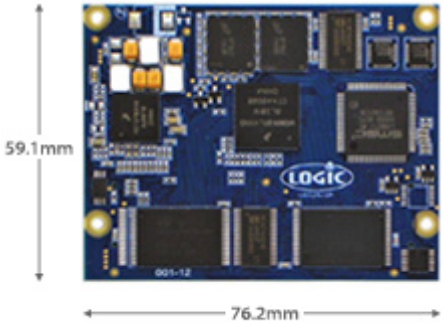
NETFLIX

Freescale i.MX31 System on a Chip

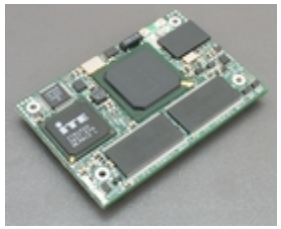


NETFLIX™

Commercial Millicomputer Module Designs



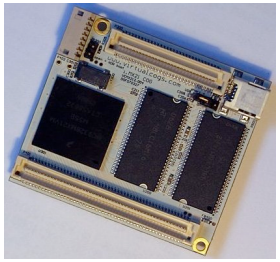
Freescale SoM 76x59mm i.MX31



Compulab 68x58mm PXA270

Colibri 68x37mm PXA320

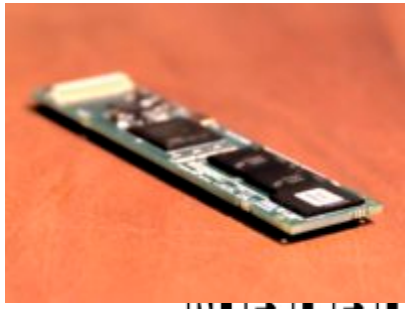
Triton 68x26mm PXA320



Virtual Cogs 50x44mm i.MX21

ADELAIDE 85x54mm i.MX31

Gumstix 80x20mm PXA270



(Most of these support up to 128MB RAM and cost ~$100)

NETFLIX™

Open Millicomputers

- Gumstix – PXA270 based schematics and PCB layout for large range of IO device modules provided at www.gumstix.com under creative commons license

- Gumstix Goliath - Embeddable Open Phone Motherboard
  - GSM/GPS/Touchscreen/Linux 2.6
  - Currently porting OpenEmbedded and OpenMoko Applications

- OPiuM – custom i.MX31 based design with 256MB RAM
  - Specified as open hardware for use by Silicon Valley Homebrew Mobile Phone Club - slow progress so far…

- Benchmarks
  - http://docwiki.gumstix.org/Benchmarks

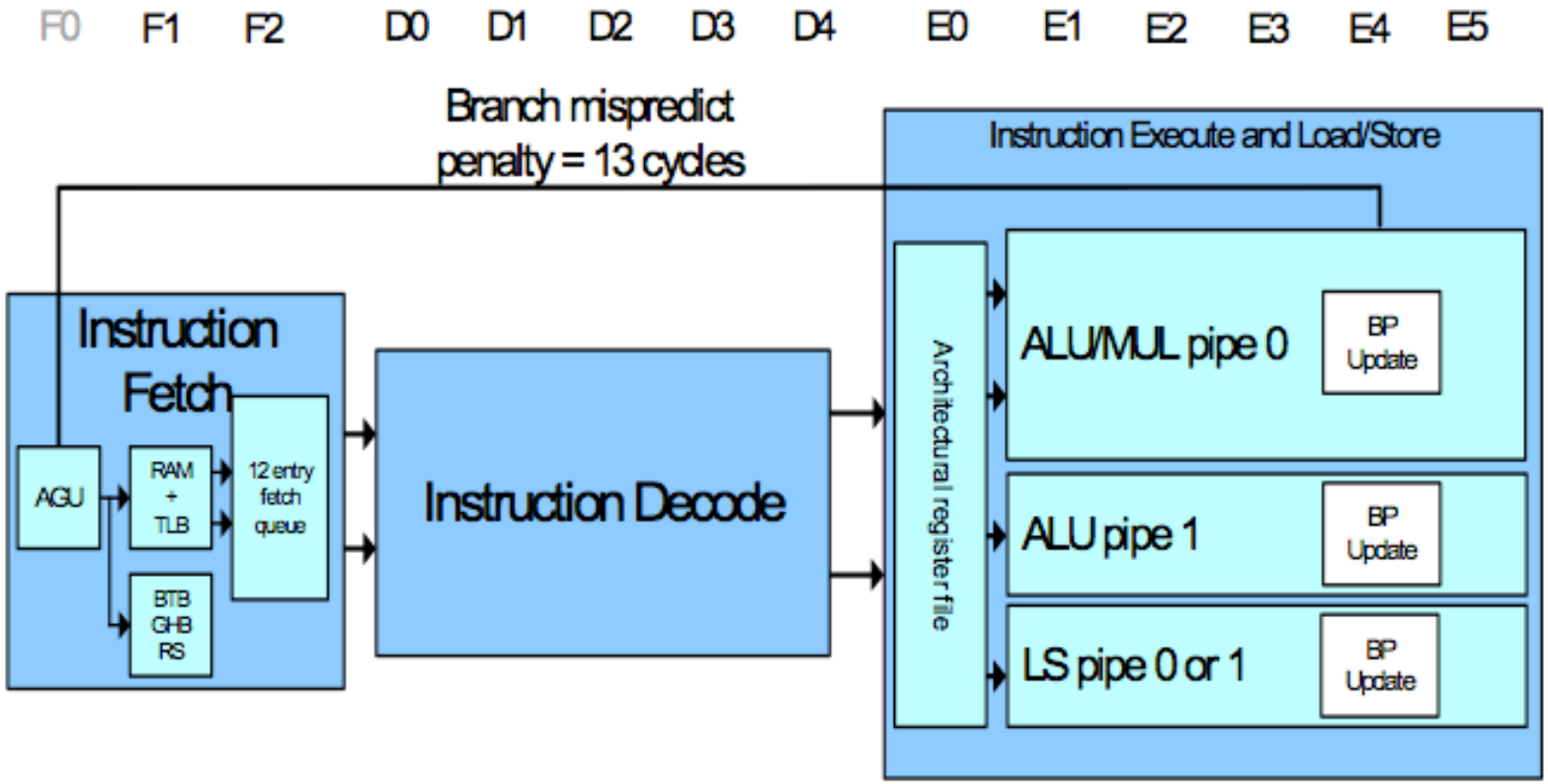
NETFLIX™

Roadmap

- All System on a Chip Designs under 250mW
- 2007 ARM Single Issue
  - 534-624MHz Common, 806MHz
  - iPhone is Samsung ~600MHz 128MB RAM
- 2008 ARM SuperScalar Cortex A8
  - Qualcomm Scorpion 1GHz+
- 2009 Intel x86 based competition?
- 2010 ARM Four Core SuperScalar Cortex A9
  - Performance claim “8x iPhone”

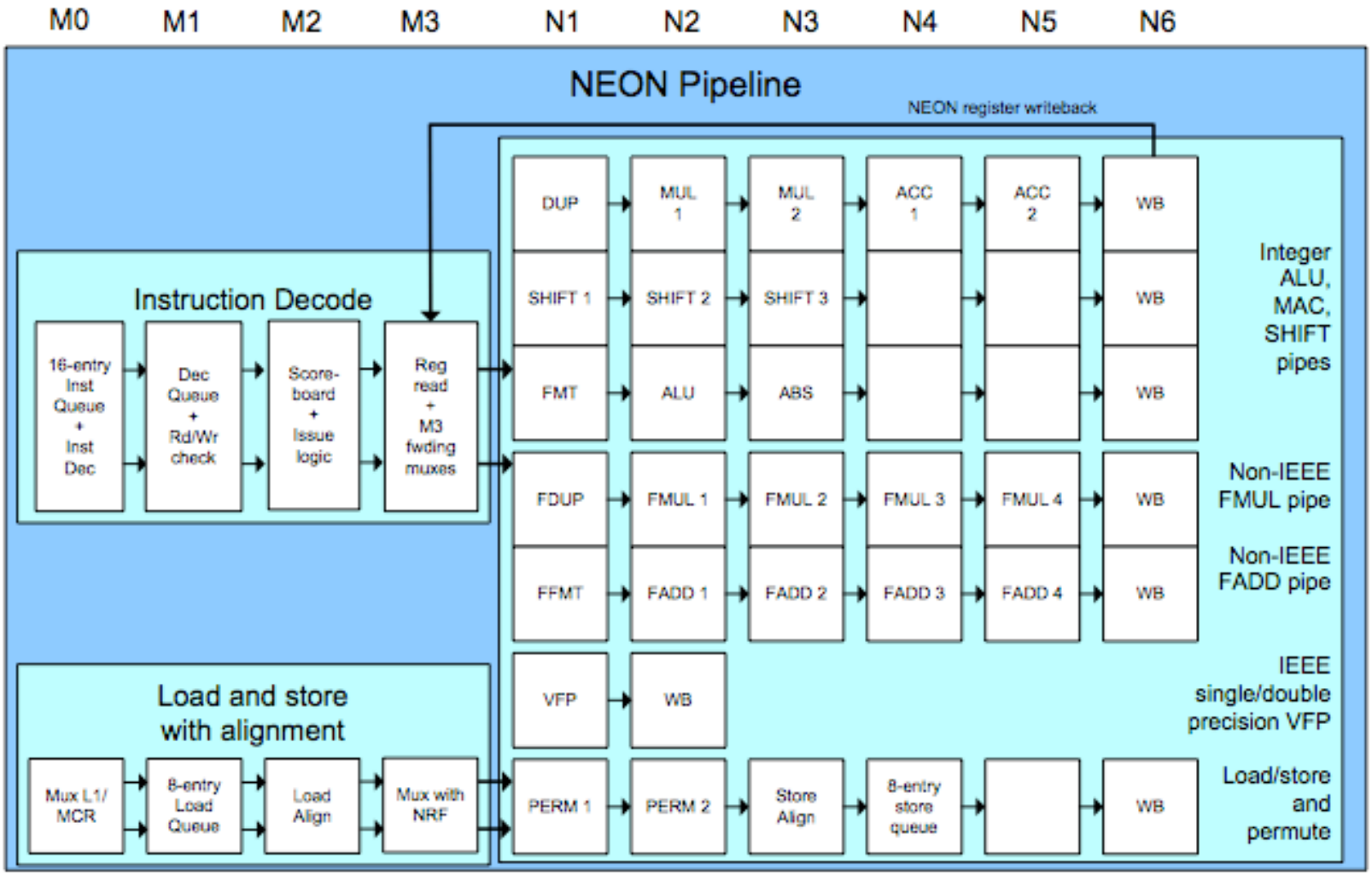
NETFLIX™

Cortex A8 Pipeline



NETFLIX

Cortex A8 NEON Accelerator



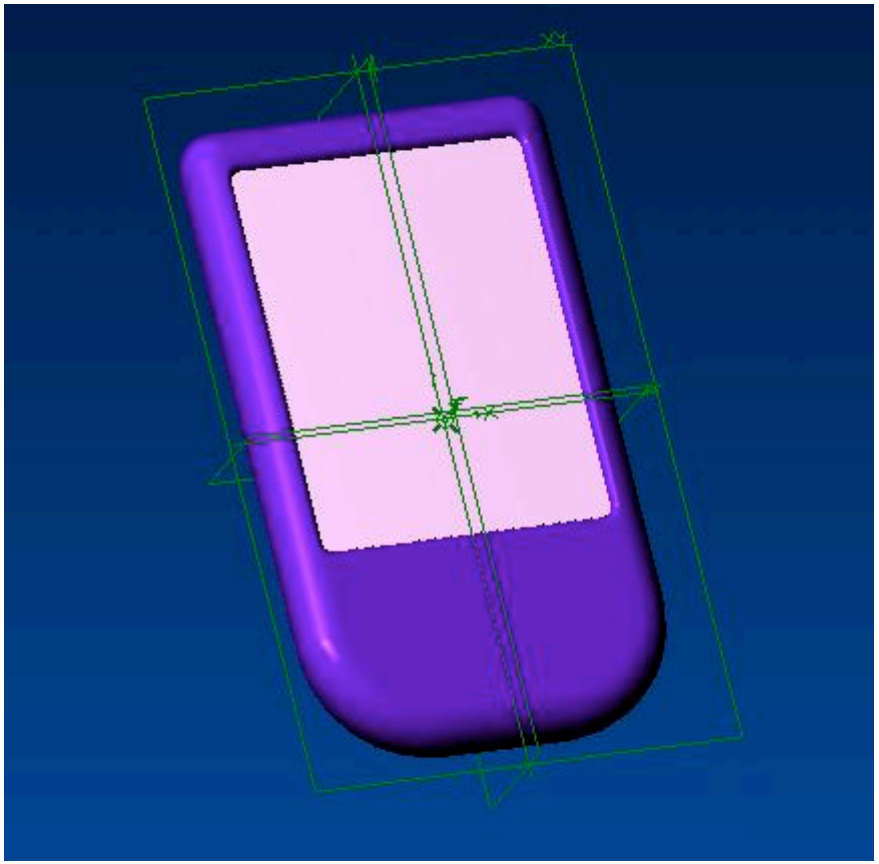
NETFLIX

Design Our Own Millicomputer

- The following presentation is speculative
  - It describes an Enterprise Millicomputer Architecture
  - A few people HomeBrewing in their spare time
  - No mainstream vendors are involved (or needed?)
  - Design specifications are subject to change
  - We may end up building nothing or a completely different design!
  - No timescales or commitments
- Additional Motivation?
  - Reduce global warming by accelerating move to millicomputing

NETFLIX

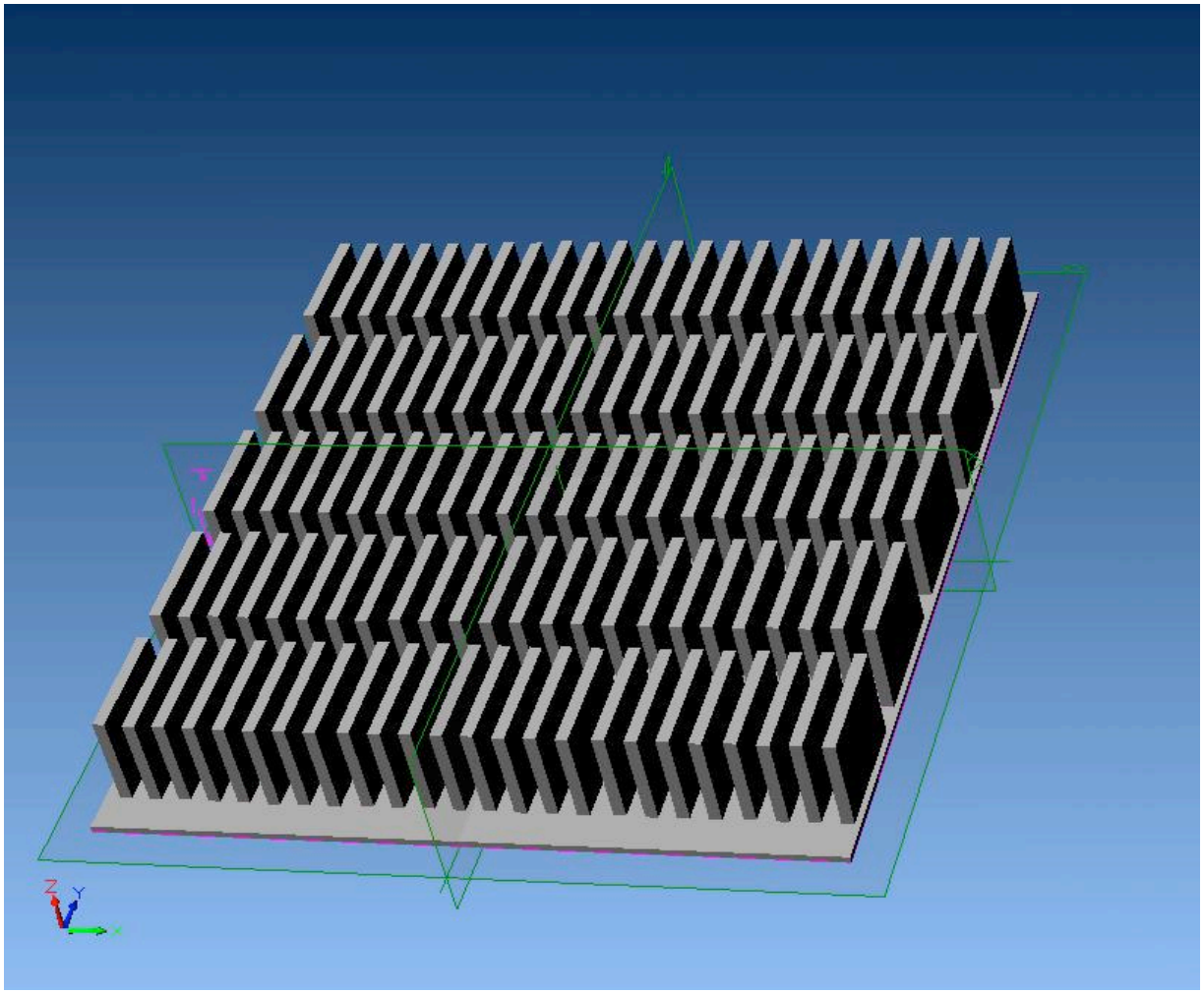
myPhone Mobile Millicomputer Packaging



CAD design shared under creative commons on Homebrew Mobile site
ABS plastic case manufactured one-at-a-time using 3D Printer
Gumstix millicomputer module mounted on phone-specific I/O PCB
Gumstix has built working homebrew phone prototypes

NETFLIX™

Enterprise Millicomputer Vertical Packaging

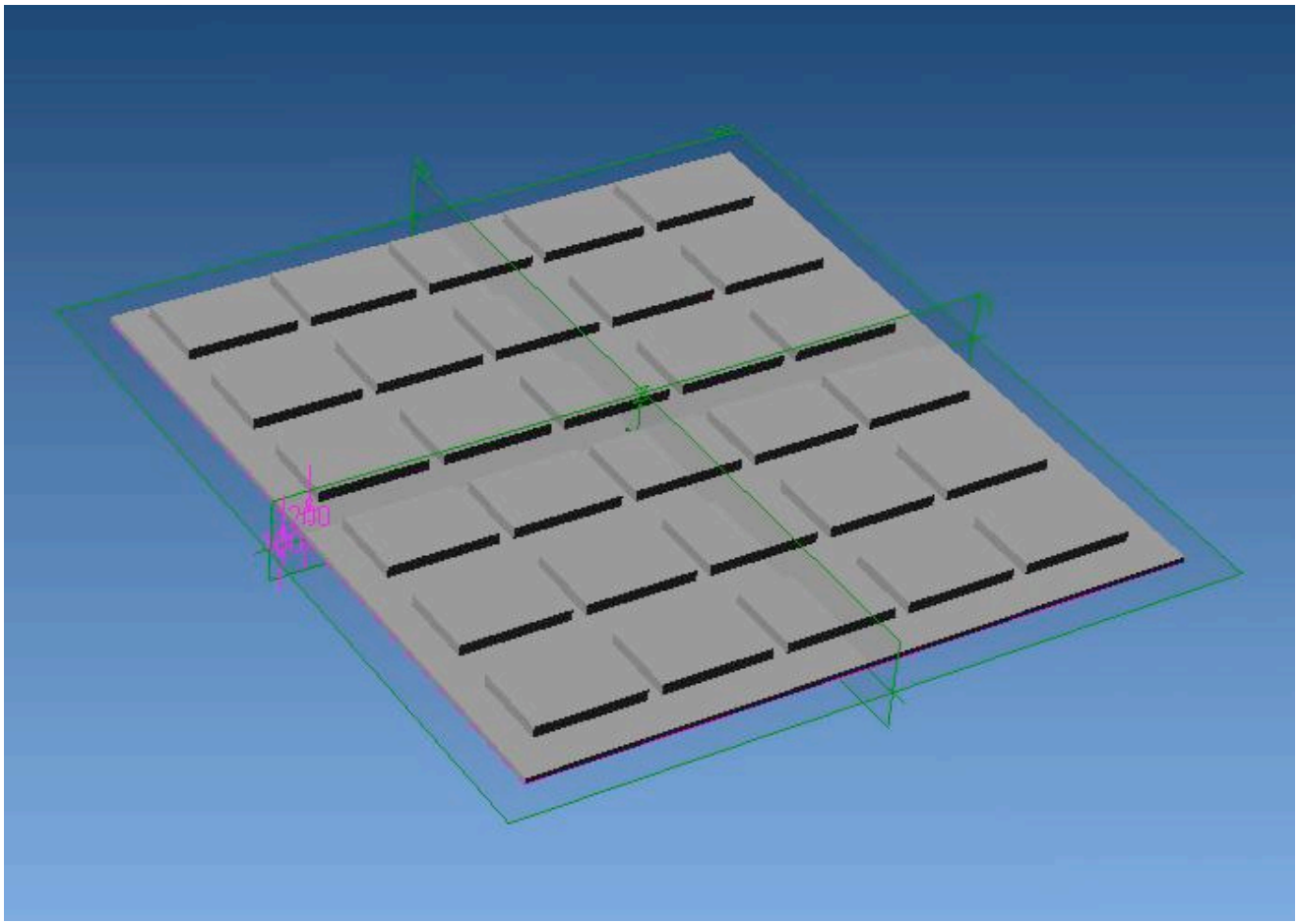


Example 1U Server package

5x24 array of modules the same volume as standard 1U enterprise motherboard

NETFLIX™

Enterprise Millicomputer Horizontal Packaging



Thin stackable 5x6 array of modules same shape as 1U enterprise motherboard

NETFLIX

Networking

- Ethernet required for external connections
  - Power draw about 1W per 1 Gbit ethernet port
  - More than the CPU, too much per module
  - Configure an Ethernet gateway per cluster
  - Implement load-balancer functions in gateway
- i.MX31 has on-chip USB 2.0 480Mbit/s
  - Use USBNet transport to route to Ethernet
  - Use High speed, low power 8-port 480Mbit/s USB switches

NETFLIX™

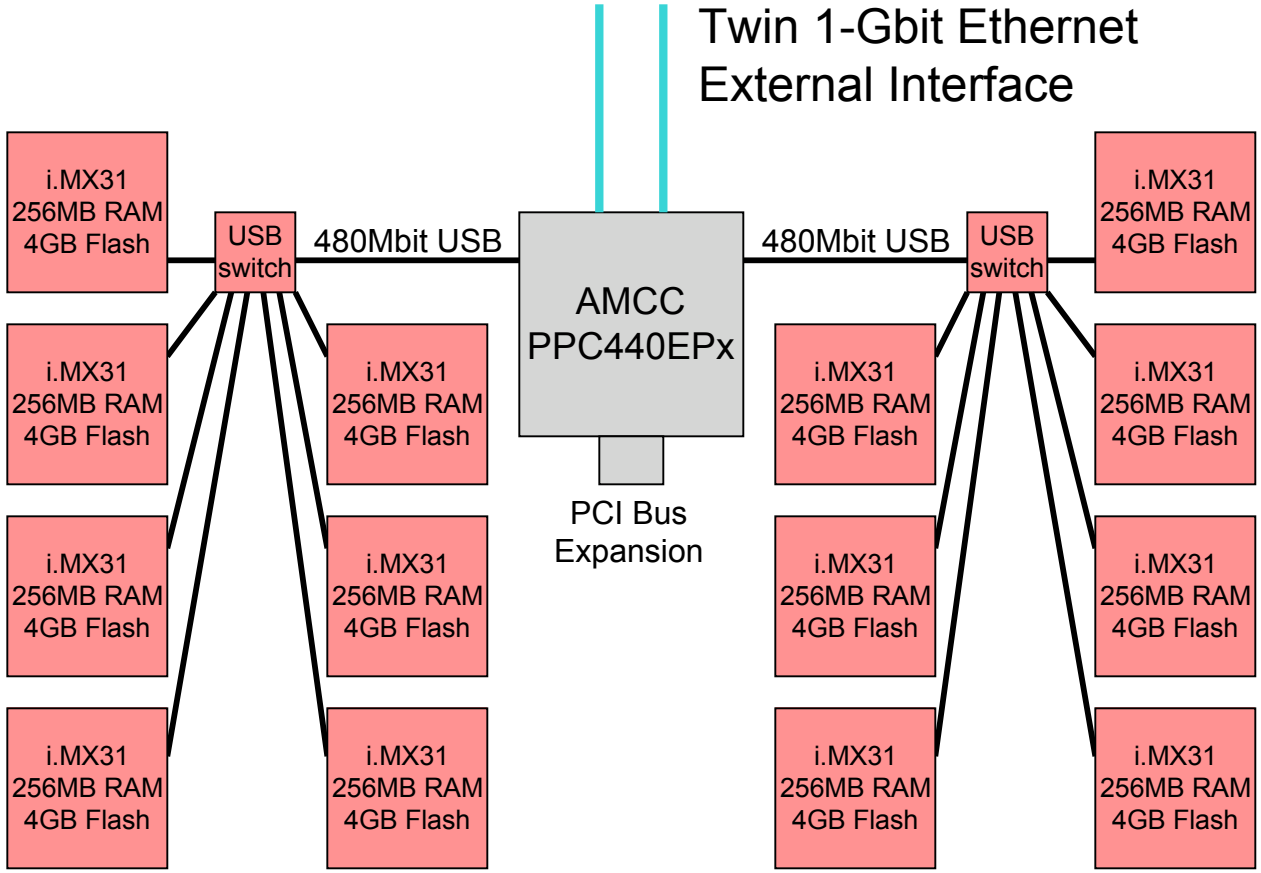
Gateway and Load Balancer

- Needs Ethernet, USB, perhaps PCI interfaces
- A few watts needed to drive Gbit Ethernet
- AMCC PPC440EPx
  - 400MHz PowerPC system on a chip @ 3W
  - Dual 1Gbit Ethernets
  - Dual 480Mbit USB2.0
  - PCI Bus Interface
- Linux Load Balancer Open Source choices
  - Haproxy, XLB, Balance, Ultra Monkey 3, vrrpd

NETFLIX

Enterprise MilliCluster

14 OPiuM Millicomputer modules behind Ethernet Bridge/Load Balancer
1 Gbit/sec redundant network, 7.5 GHz CPU, 3.5 GBytes RAM, 56 Gbytes Storage
5.5” Wide x 12” Deep x 0.4” High - 3 Watts Idle, 20 Watts Peak, no heat sinks!

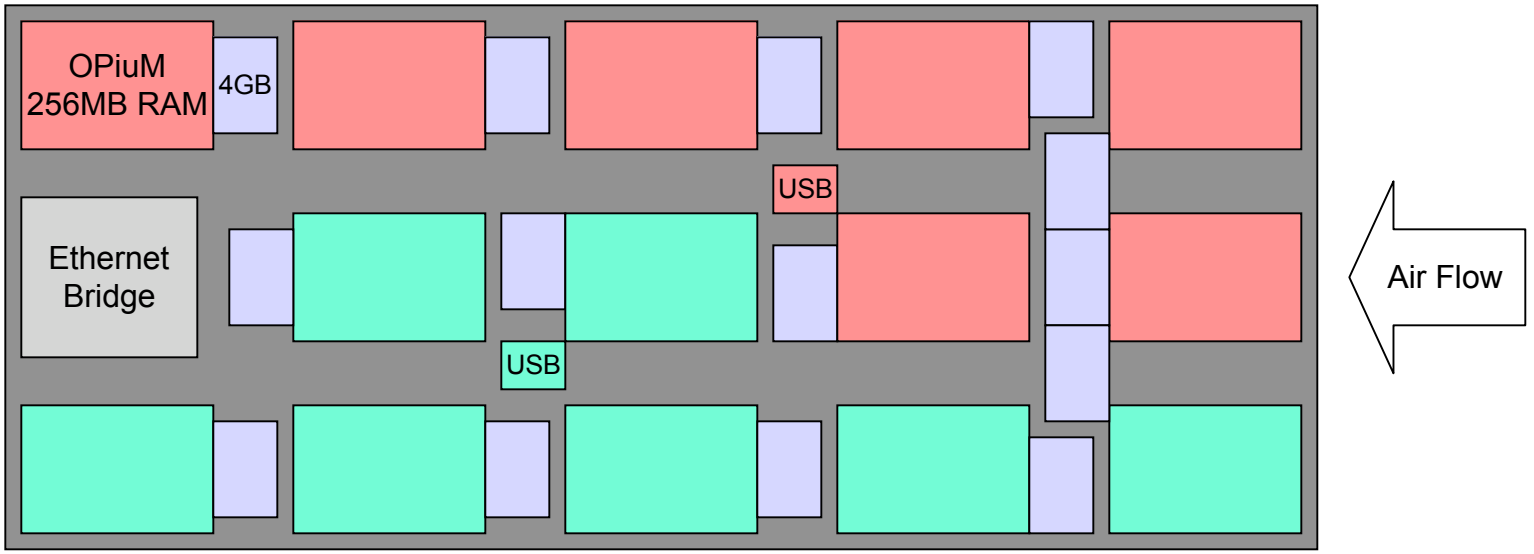


NETFLIX

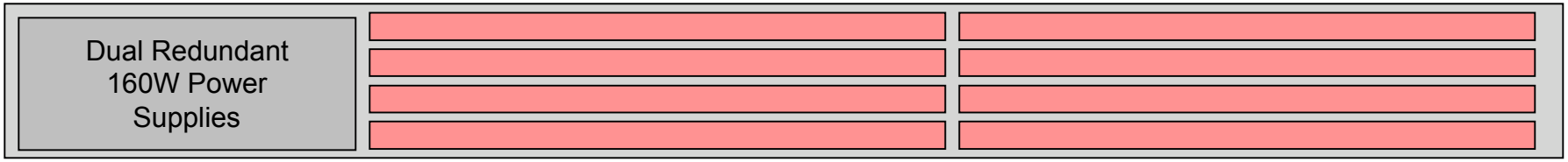
Enterprise MilliCluster Packaging

Stack Side by Side Four Deep in 1U Package

Top view of One MilliCluster. Ethernet Bridge at Rear of Package
14 OPiuM I.MX31 Modules and microSD card mounts



Cross Section Through 1U Package Showing Eight MilliClusters, Rear Panel Has 16 x 1 Gbit Ethernet Ports



NETFLIX

Enterprise Millicomputer Spec Overview

- Standard 1U Enterprise Server Package contains Eight MilliClusters
- Density – 112 OPiuM modules per RU, 4704 modules in 42RU rack
- Power – Peak 160 Watts/RU, Idle 24 Watts/RU, Peak ~6.7KW/Rack
- CPU - Performance total 60 GHz/RU, 28 GBytes/RU RAM
- Network - 8 Load balancer/bridge-routers per RU
  - 8 Gbits/RU module bandwidth on 16 redundant Gbit ports
  - Ethernet switch could be added to design to reduce port count
- Storage – microSD flash memory socket at each module
  - 2 GB microSD for very low cost, 4 GB for capacity, 8 GB in 2008
- Optional Extras
  - Disk – modules all include ATA disk controller if needed
  - Graphics – i.MX31 modules include OpenGL 3D graphics engine
  - Display – modules all include LCD display driver, touch screen
  - I/O – modules include multiple USB/serial interfaces etc.

NETFLIX

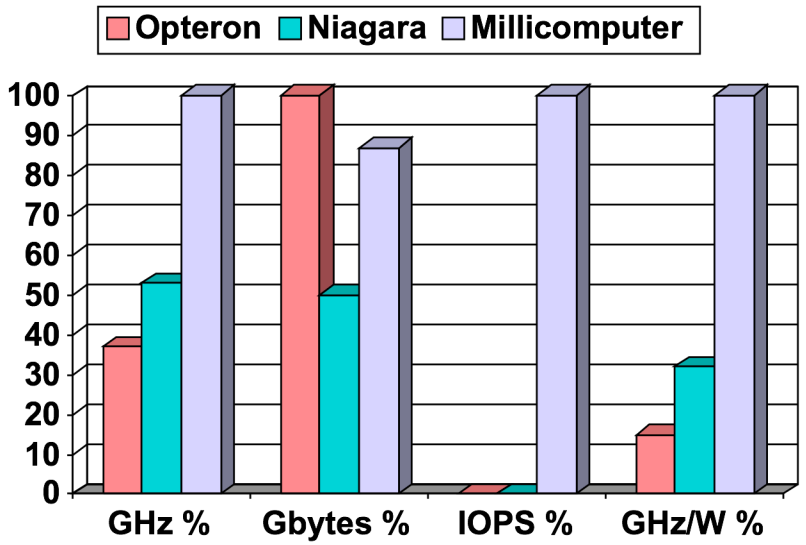
The Flashiest Storage

- Per-module Flash in microSDHC format
  - http://www.getflashmemory.info/category/microsdhc/
  - 2 - 8 GBytes of NAND Flash in one microSDHC
  - Streaming read and write performance ~20MByte/s
  - NO SEEK! Random access at 1000's of IOPS
  - 112 x 4 GB = 448 Gbytes/RU, 18.8TB/Rack
  - 112 x 20 MB/s = 2240 MB/s/RU, 94 GB/s/Rack
- Per-module "Spinning Rust" Disks?
  - One ATA disk interface connector per module
  - Route one module per MilliCluster to connector
  - Connect four MilliClusters to disks or larger SSDs

NETFLIX

Packaging Comparisons in 1U

- Sun x4100 Opteron 1U - 400W
  - CPU performance is probably double at the same GHz
  - 2.8 GHz four cores x 2 = 22.4 GHz, 32 GB RAM - $13K list
- Sun T1000 Niagara 1U – 200W
  - 1.0 GHz 8 core/32 threads = 32 GHz (optimistic), 16 GB RAM - $15K list
- Enterprise Millicomputer 1U – OPiuM i.MX31 based – 160W
  - 532 MHz x 112 = 60 GHz, 28GB RAM - $14K?
- Millicomputer Networking
  - Higher Network bandwidth
  - No external Load balancer
- Millicomputer Storage is No Contest!
  - 2x146GB disks 240 IOPS vs. ~500000 IOPS, 448 GB Flash



Prices from www.sun.com June 2007

Actual Performance benchmarks still need to be measured!

NETFLIX

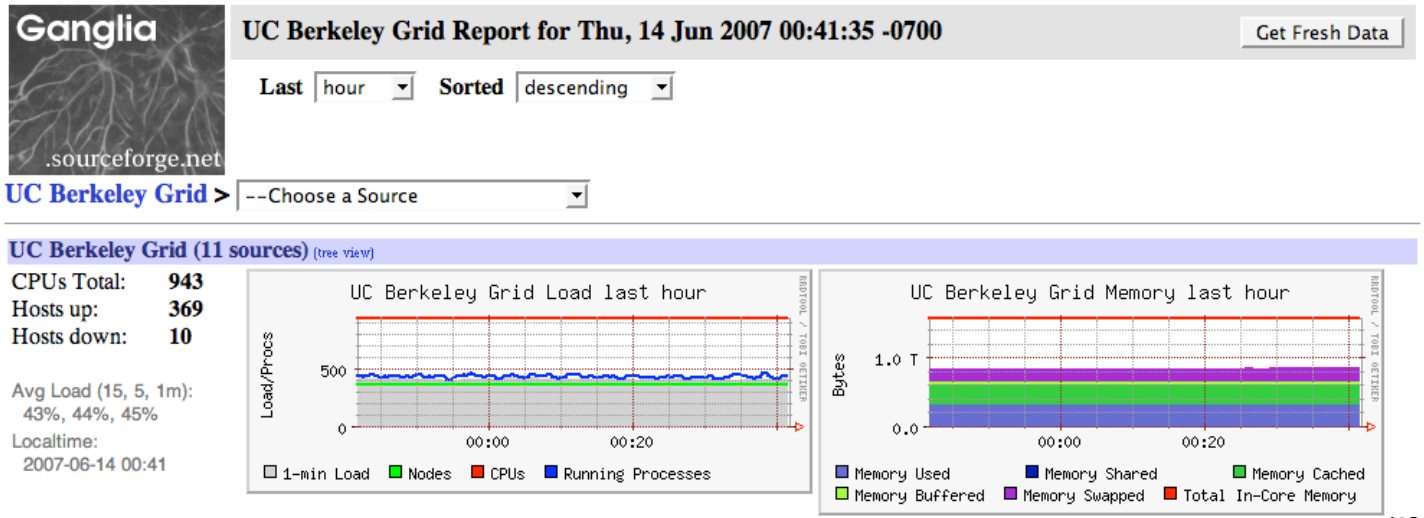
Software Implications

- Application memory size limit: 128-256MB
- Similar to mainstream systems from 2001
  - Sizes will catch up eventually
    - Classic “disruptive innovation” characteristic
  - Suitable for web content delivery
    - Especially intense random storage access
    - Static content and cache servers
    - Horizontally scaled MySQL services
    - Small Java applications: Hadoop? KETL?
  - Video wall “Cave” display driver
    - Use i.MX31 to drive tiled video outputs

NETFLIX™

Management Implications

- Large scale “grid” of small systems
  - Needs lightweight monitoring agent
  - Needs aggregation tools
  - Load balancer integration/awareness
- Ganglia? http://ganglia.sourceforge.net/



NETFLIX™

Summary

- For Similar 1U package, similar cost per package
  - Power less than a Niagara, less than half an Opteron system
  - Total RAM capacity similar
  - Raw CPU GHz double, GHz per Watt five times Opteron
  - Flash storage is 1000x faster for both random/sequential IOPS
- Applications that can be broken into small chunks
  - Small scale or horizontally scalable web workloads
  - Legacy applications that used to run on 5 year old machines
  - Graphical video walls and storage I/O intensive applications

NETFLIX™

Next steps

Performance and power benchmarking and validation

Seek out collaborators

acockcroft@netflix.com

http://www.millicomputing.com

Build prototypes…

Save Power!

NETFLIX