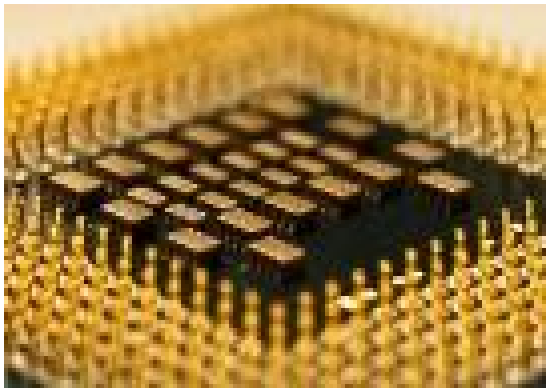
漫谈面向大数据,云计算平台建设的新视角与新技术

Prof. Cho-Li Wang
The University of Hong Kong

Outline

- **(1) Crocodiles Project**:
  - **Big Data Computing on Future Maycore Chips**
    - **(1) Kilo-Sim; (2) Rhymes (Software CC on Intel SCC)**
- **(2) 异构多核系统优化与并行计算 (863 project)**
  - Java Parallelization on GPUs and Intel MIC.
- **(3) Operating System for Future Data Center**
  - **OS-1K (马其顿方阵) 解耦操作系统**

(1) *Crocodiles Project*: Cloud Runtime with Object Coherence On Dynamic tILES for future 1000-core tiled processors" (1/2013-12/2015, HK GRF)



Big Data Computing on Future Maycore Chips

CPU Designers Debate Multi-core Future (2008)

- Chuck Moore, an AMD senior fellow
  - Heterogeneous collections of cores.
    - By 2018: 10-teraflop x86 APU computing node
  - 2014: APU (Fusion, 2011) → 'Berlin' Opteron (4/2014, HSA)

- Anant Agarwal, founder of startup Tilera, MIT:
  - Homogeneous collections of general-purpose cores
    - By 2017: embedded (4,096 cores); server (512 cores).
  - 2014: Tilera TILE-Gx8072 (72-core, 3/2013), @1~1.2GHz.

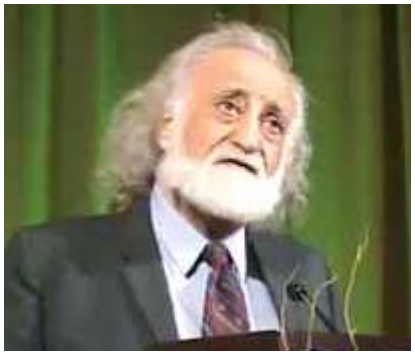
- Rick Hetherington, Oracle
  - "Thread-rich" cores (e.g., 500~1,000 threads/core).
    - By 2018: servers may have only 32-128 cores.
  - 2014: SPARC T5 (2013) - 128 threads (28 nm, 16 cores x 8 threads/core).

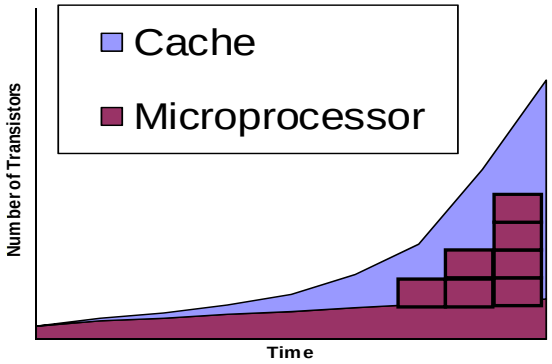
- Shekhar Borkar, director of Intel's Extreme-scale Technologies
  - Microprocessor cores will get increasingly simple.
    - Intel SCC could theoretically scale to 1,000 cores*
  - 2014: Intel 60-core Xeon Phi → Knights Landing (72 Atom cores)

* http://www.zdnet.com/intel-why-a-1000-core-chip-is-feasible-3040090968/

CPU Designers Debate Multi-core Future (2008)

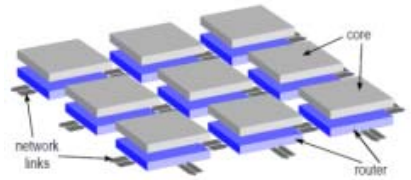
- **Yale Patt**, a well-known computer architect, UT Austin, National Academy of Engineering (2014):
  - **Multi-core is just "multi-nonsense"**
    - **"multi-core is a solution looking for a problem."** → *"was the easy way out chosen by Intel/AMD to use up the increasing transistors"* (by adding more cores & cache).





Multi-core → *Multi-opportunity?*

The 1000-Core Era



Multicore system simulated by Hornet. Source MIT

- Experts predict that by the end of the decade we could have as many as 1000 cores on a single die (***S. Borkar, "Thousand core chips: a technology perspective"*** **DAC 2007**).
- **International Technology Roadmap for Semiconductors (ITRS) forecast:**
  - **450 cores by 2015 → 1500 cores by 2020**
- **Why 1000-core chip ?**
  1) **Space efficiency:**
     - **densely packed servers cluster**
  2) **Power Efficiency : Greener**

Intel: Why a 1,000-core chip is feasible

Intel engineer Timothy Mattson tells how the company's 48-core Single-chip Cloud Computer could theoretically scale to 1,000 cores.

HPC wire

MIT's Hornet

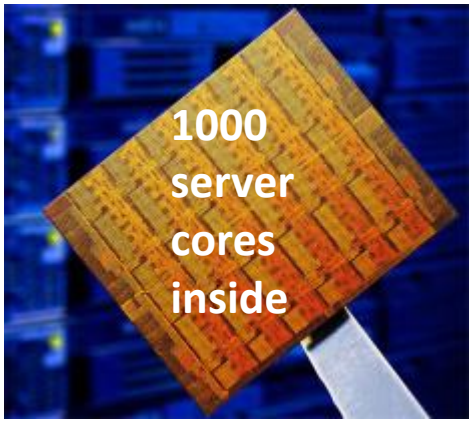
March 06, 2012

Chip Simulation for the 1000-Core Era

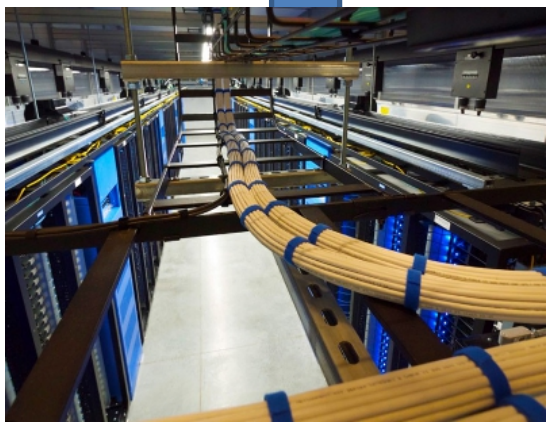
Forecasting the Future Datacenter: "Data Center on a Chip" ?

Microsoft's Chicago Data Center
(2,000 servers per container)

Data Center on a Chip?

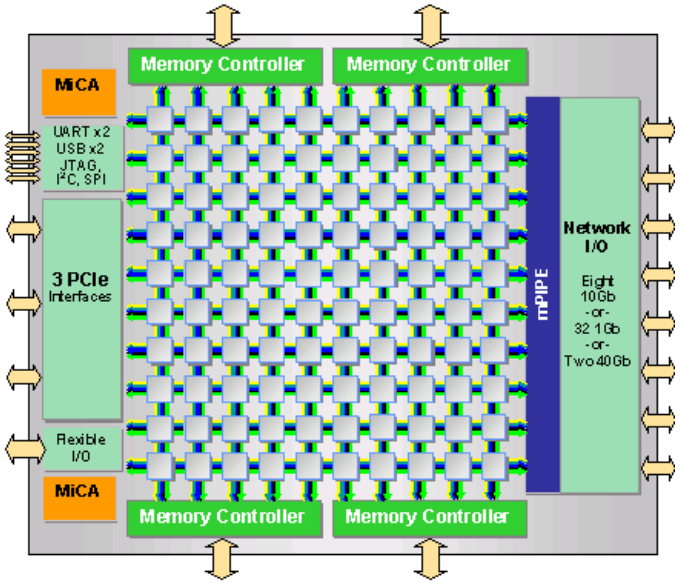


Google Data Center

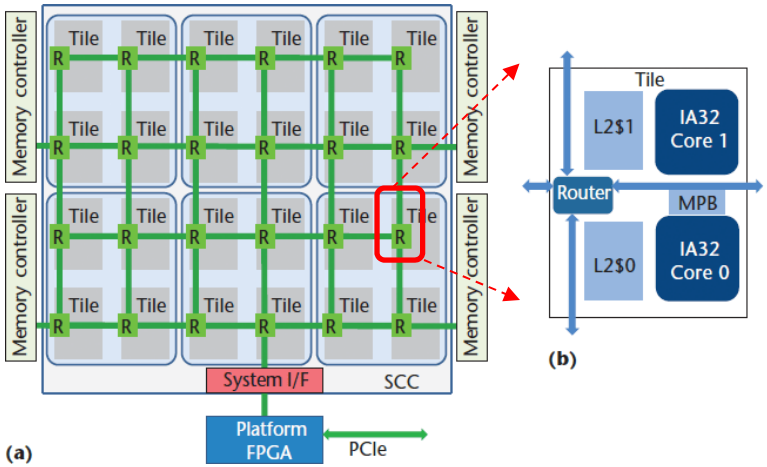


Facebook's Data Center at Prineville

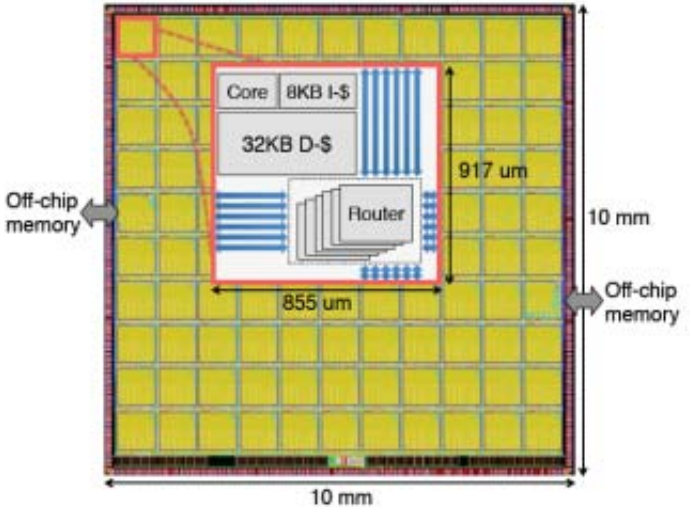
Survey of Tiled Manycore CPUs



Tilera Tile-Gx100 (100 64-bit cores)



Intel's 48-core Single-chip Cloud Computer (SCC)

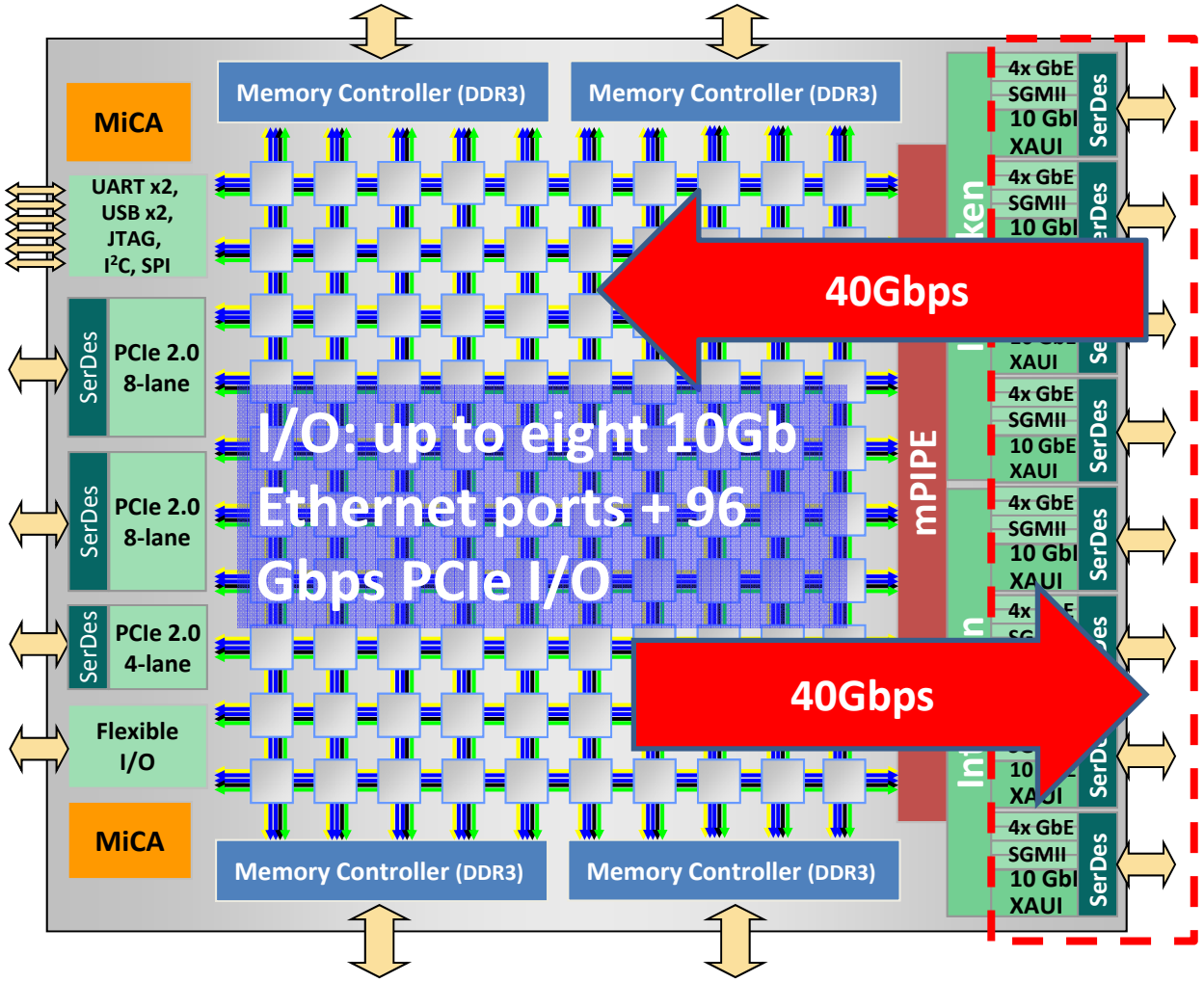


MIT's 110-core EM² chip (2013)

Intel's 14nm 2nd generation Xeon Phi (Knights Landing): 2015?

Tilera's TILE-Gx100:

Complete System-on-a-Chip with 100 64-bit cores



100 64-bit Processor Cores

- 1.0GHz -1.5GHz
- **non-x86 architecture**
- **32 MBytes** total cache
- **Full Hardware Cache Coherence**

Interconnect:

- 2D on-chip mesh network (iMesh)

DDR3 RAM: >475 Gb/sec DDR3

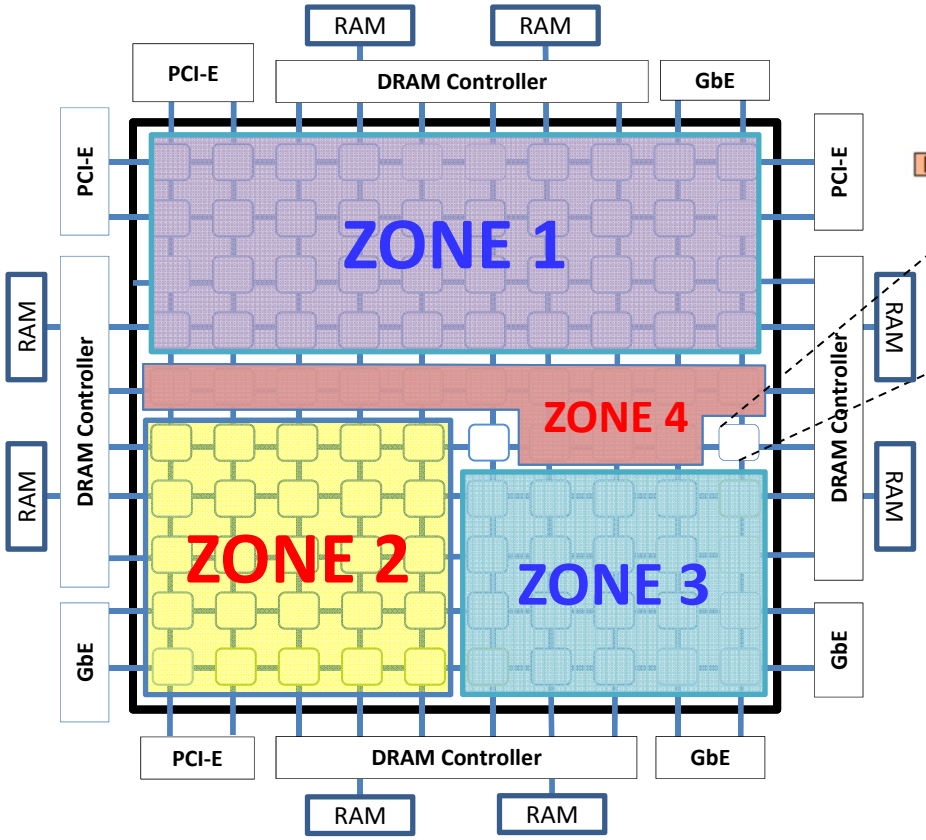
OS: Tilera Linux

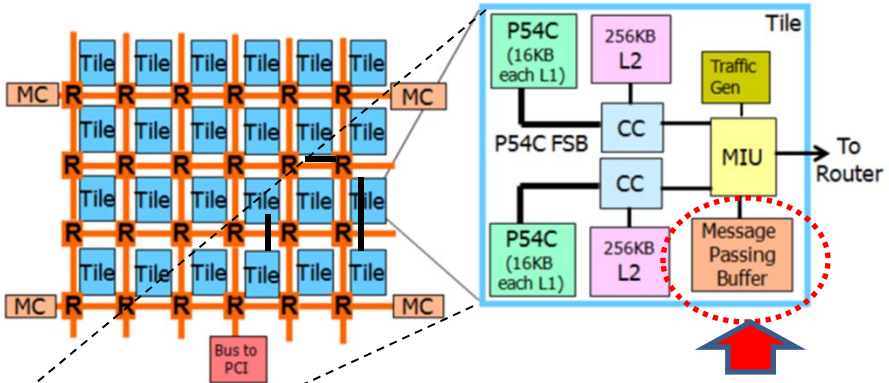
TILEncore-Gx72 (8 x 10G + TILE-IQ) → 40Gbps SDN Forwarding Plane, Open vSwitch (OVS) , deep packet inspection (DPI), … (4/2014)

Crocodiles: Cloud Runtime with Object Coherence On Dynamic tILES for future 1000-core tiled processors”

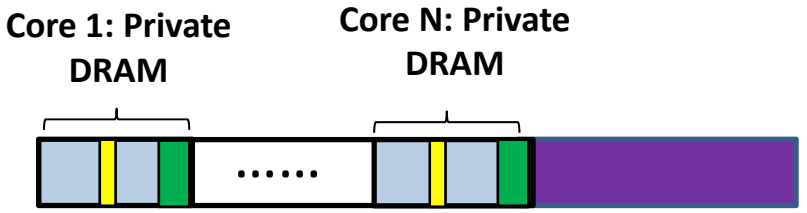
(HK GRF: 01/2013-12/2015)

Current Platform: Intel’s 48-core Single-chip Cloud Computer (SCC) + Barrelfish OS





Programmable On-chip Messaging Passing Buffer (MPB)

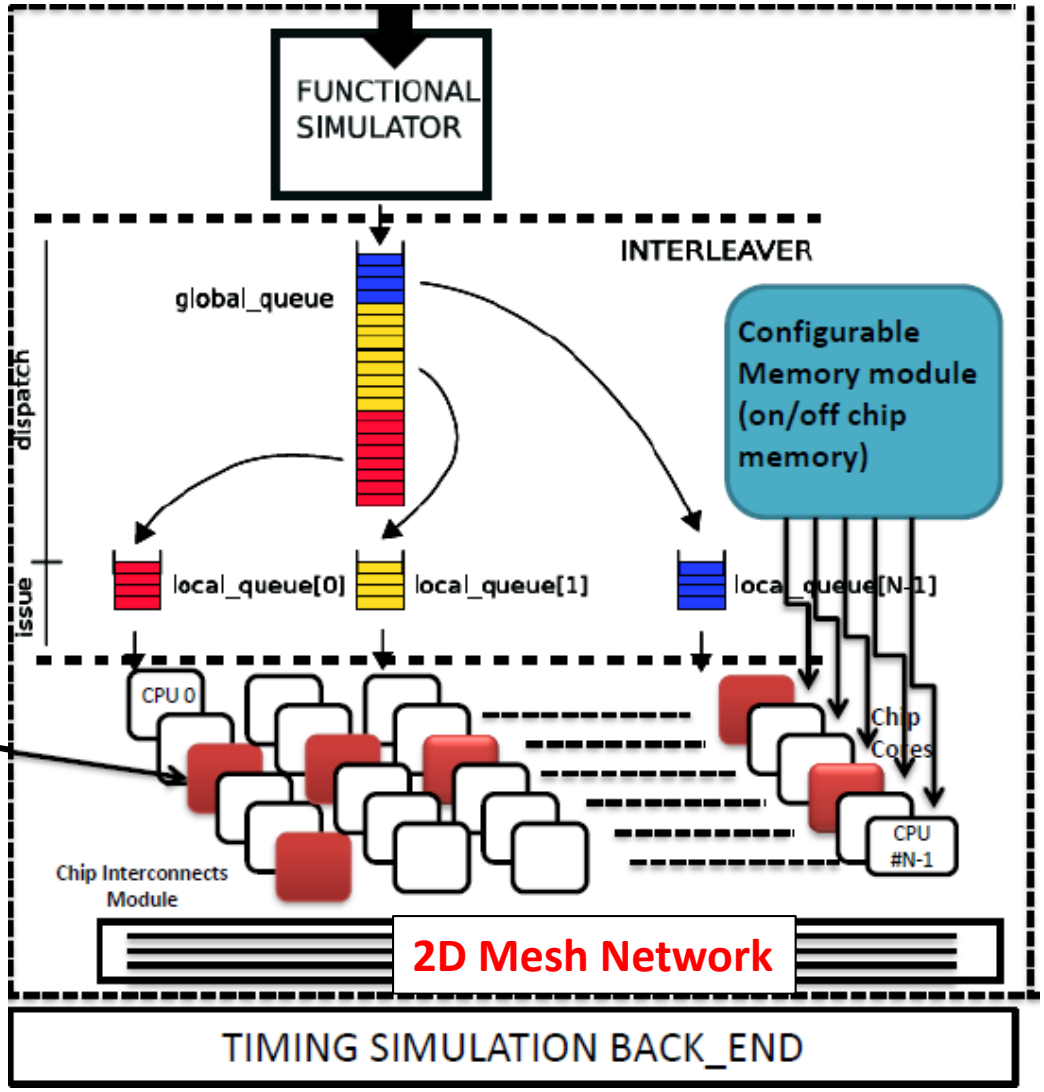


Core’s Private DRAM
Message Passing Buffer
Global Shared DRAM
OS Kernel

Multi-kernel OS

HKU Kilo-Sim (on-going)

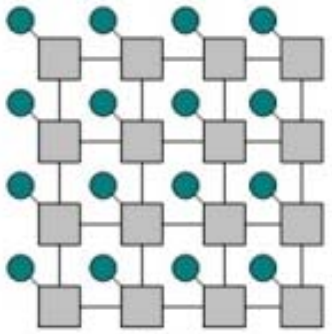
A Scalable and Cycle-accurate Full-System Simulator for Manycore Chips



McPAT: power modeling

Barrelfish OS

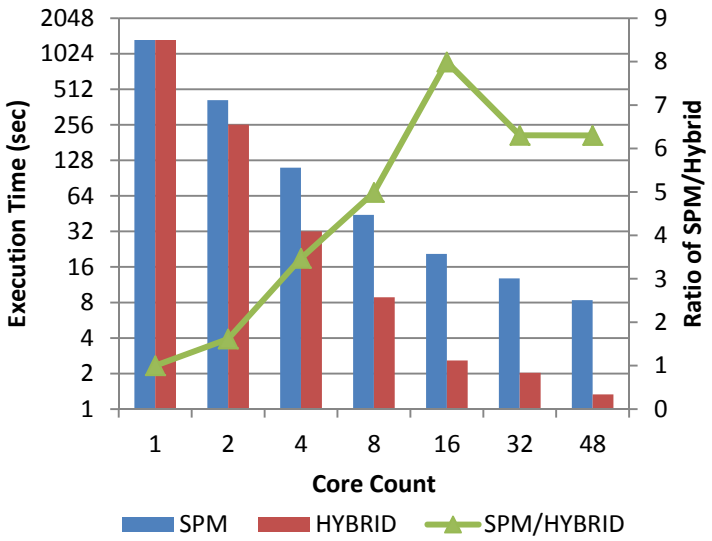
Cycle-accurate interconnection network model based on GARNET



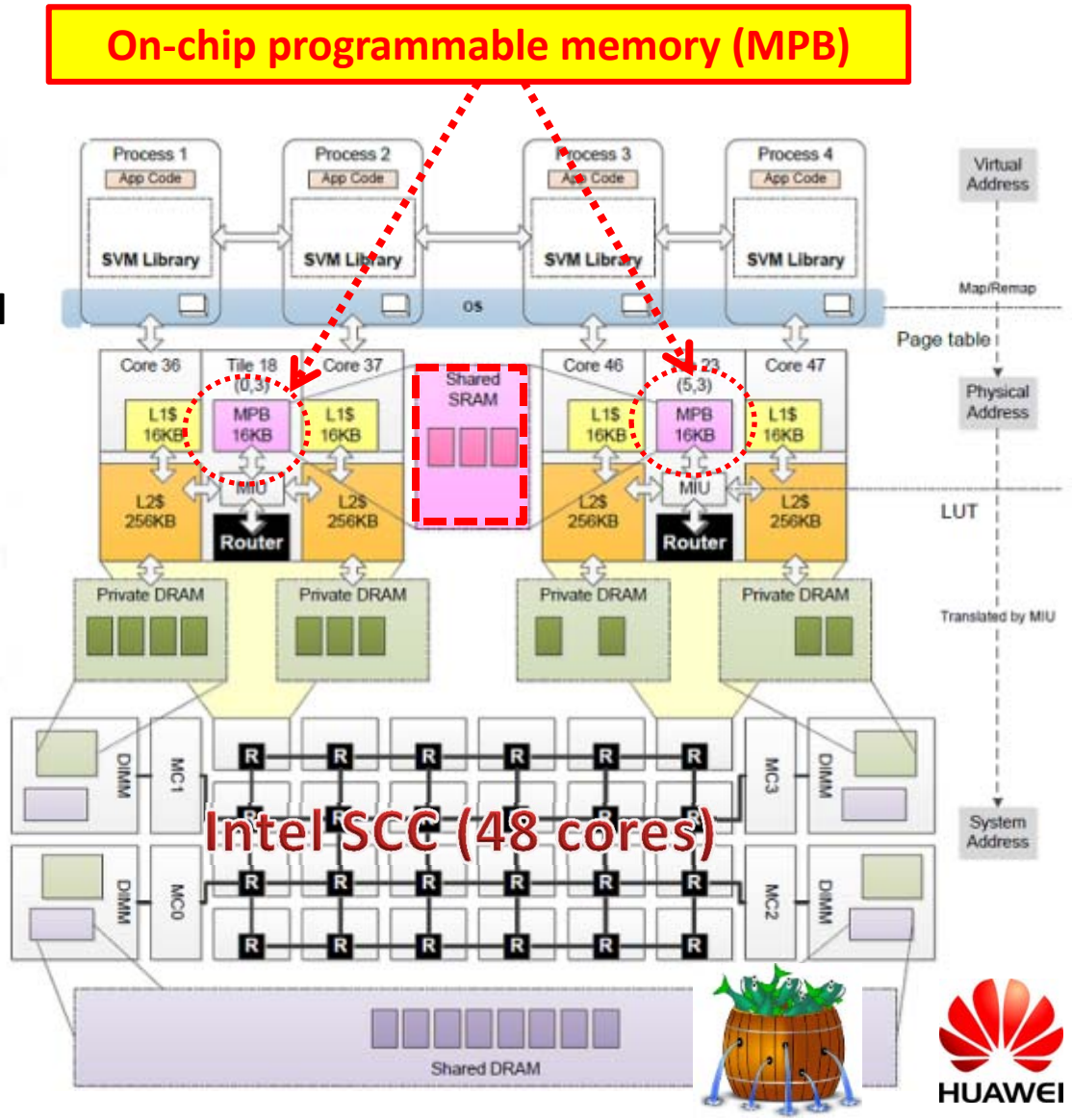
Mesh

Software-Managed Cache Coherence: Rhymes

- Barrelfish multi-kernel OS
- Leverage programmable on-chip memory (MPB)
- Scope Consistency (ScC) Model: minimizing on-chip network and off-chip DRAM traffic → 6x ~ 8x faster than Intel's SPM.



"Rhymes: A Shared Virtual Memory System for Non-Coherent Tiled Many-Core Architectures".



国家863计划

(2) 863 Project (3/2014-)

异构多核系统优化与并行计算

Other Members

Tianhe-2 (33.86 pflops) @Guangzhou (1st in Top500)

Tianhe-1A @ Tianjin (2.56 pflops , 13rd) & @ Hunan (0.77 pflops, 42th)

Nebulae (1.27 pflops) @Shenzhen (21st in Top500)

Sunway Blue Light (神威蓝光) @Jinan (0.796 pflops, 40th in Top500)

超级计算创新联盟 **(9.23. 2013)**

Top500: 06/2012

(GPU → Low Efficiency?)

Rank	Computer	Cores	R_{max} (TFlops)	R_{peak} (TFlops)	Efficiency
1	LLNL **Sequoia** – **BlueGene/Q**, USA	1572864	16,324.75	20,132.66	**0.81**
2	RIKEN **K computer,** Japan	705024	10,510.00	11,280.38	**0.93**
3	ANL **Mira, BlueGene/Q**, USA	786432	8,162.38	10,066.33	**0.81**
4	**SuperMUC**, IBM iDataplex, Germany	147456	2,897.00	3,185.05	**0.90**
5	**Tianhe-1A,** China	186368	2,566.00	4,701.00	**0.54**
6	ORN **Jaguar**, USA	298592	1,941.00	2,627.61	**0.73**
7	**BlueGene/Q**, Italy	163840	1,725.49	2,097.15	**0.82**
8	**JuQUEEN**, **BlueGene/Q**, Germany	131072	1,380.39	1,677.72	**0.82**
9	**Bull**, France	77184	1,359.00	1,667.17	**0.81**
10	**Nebulae (星云)**, China	120640	1,271.00	2,984.30	**0.42**

天河: **7168 NVIDIA M2050 GPUs**
星云: **4640 NVIDIA C2050 GPU**

#1 in Top500 (11/2013):
Tianhe-2 Supercomputer

- 16 000 nodes (3.12 million cores), each with
  - Two 12-core Intel Xeon IvyBridge processors @ 2.2 GHz
  - Three 57-core Xeon Phi processors @ 1.1 GHz
  - via **PCI-E 2.0** ← **Bottleneck !**
  - 64 GB DDR3 memory + each Xeon Phi has 8 GB per Xeon Phi → Total **88 GB** of memory per node
  - 3.432 Tflops per node
  - **CPU : MIC = 1: 7.1**
- **整体总计内存 (total memory): 1.404 PB**
- **国防科技大学研制**
- 國家超級計算廣州中心

Linpack (Rmax)	33.8 PFlop/s
峰值速度 (Rpeak)	54.9 PFlop/s

LOW Efficiency: 61.7%

Heavily rely on MICs! But still low efficiency

#2 in Top500 (11/2013):
Titan @ Oak Ridge National Lab.

LOW Efficiency: 64%

- 18,688 AMD Opteron 6274 16-core CPUs (32GB DDR3).
- 18,688 Nvidia Tesla K20X GPUs
- Total RAM size: over 710 TB
- Total Storage: 10 PB.
- Peak Performance: 27 Petaflop/s
  - GPU: CPU = 1.311 TF/s: 0.141 TF/s = 9.3 : 1
- Linpack: 17.59 Petaflop/s
- Power Consumption: 8.2 MW

Heavily rely on GPUs !

Titan compute board: 4 AMD Opteron + 4 NVIDIA Tesla K20X GPUs

NVIDIA Tesla K20X (Kepler GK110)
GPU: **2688** CUDA cores

(2) 863 Project (3/2014-)

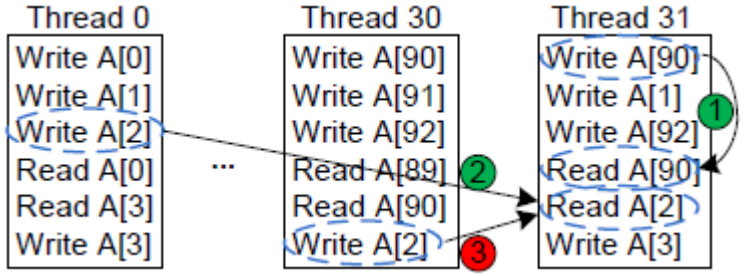
国家863计划

异构多核系统优化与并行计算

(1) JAPONICA : Java with Auto-Parallelization ON GraphIcs Coprocessing Architecture

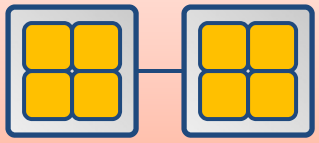
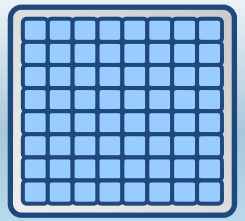
(2) GPU-TLS : Thread-level Speculation on GPU

- Assign the tasks among CPU & GPU according to their ***dependency density (DD)***



Dynamic Profiling → Scheduler → Dependency density

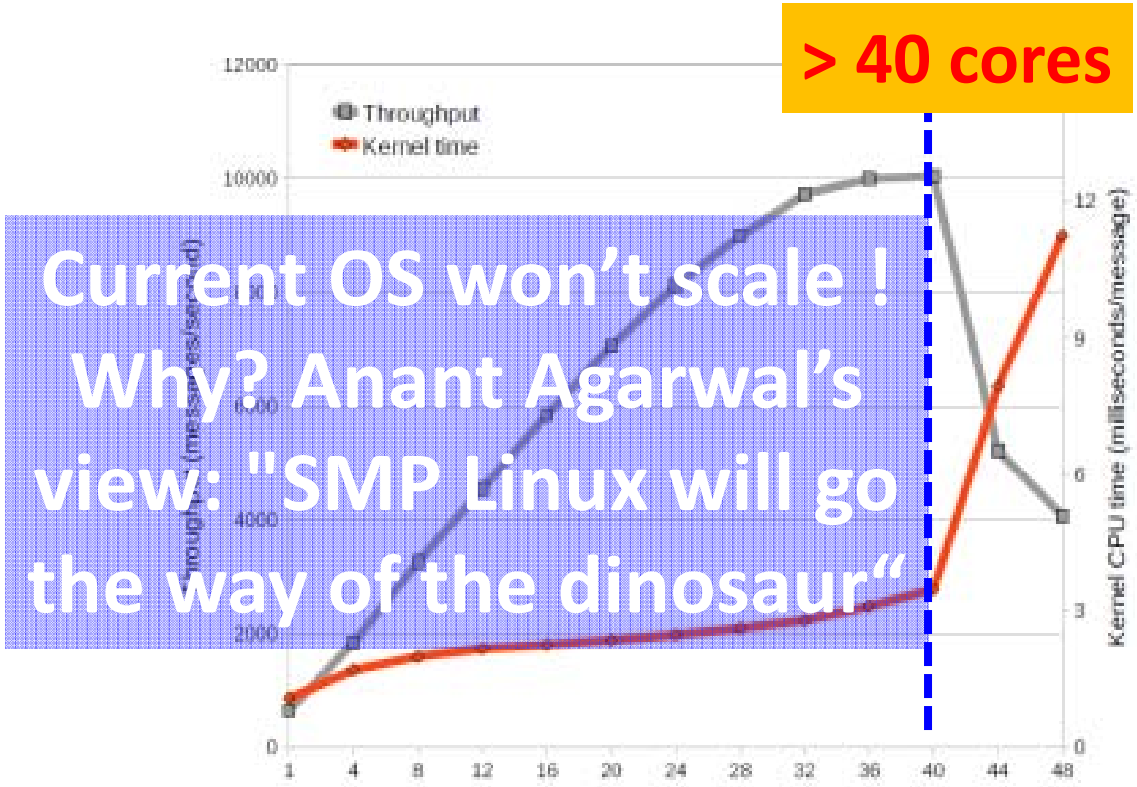
Inter-iteration dependence:
-- Read-After-Write (RAW)
-- Write-After-Read (WAR)
-- Write-After-Write (WAW)

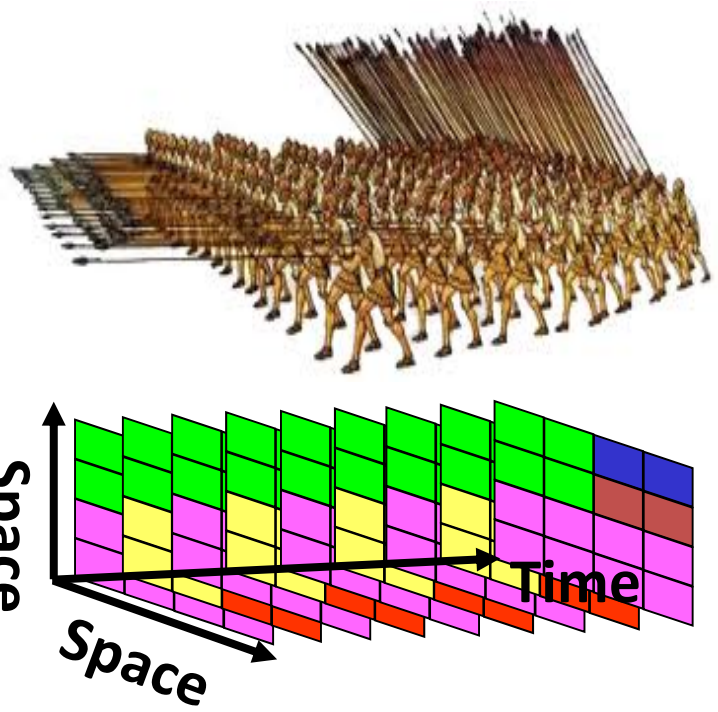
- **High** → **Parallel**
  - 
  - x86/ARM multicores
  - Multi-core CPU
- **Medium** → **Highly parallel**
  - 
  - 
  - 61 x86 cores (Intel MIC)
- **Low/None** → **Massively parallel**
  - 
  - 

  - 2880-cores GPUs (Nvidia K40)

(3) OS-1K (马其顿方阵) 解耦操作系统

- **Next-generation Cloud OS for 1000-core processor**





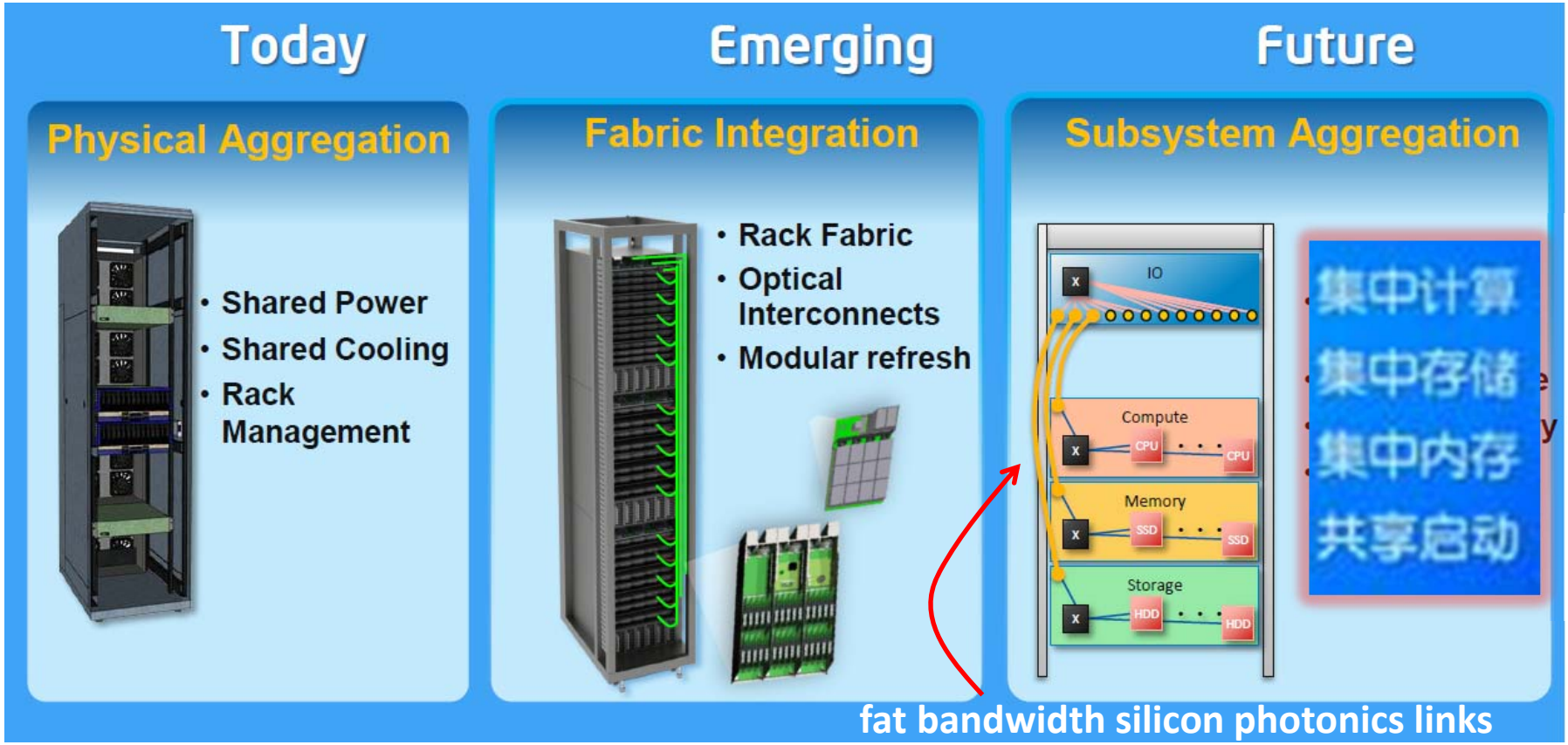
"NoHype" + "Disaggregated" Cloud OS Design:

(1) "Dynamic Zone Scaling": partitioning varies over time

(2) Dynamic clustering with "islands of cache coherence"

Intel Rack Scale Architecture (RSA)

- **High-speed, fat bandwidth silicon photonics links** to glue processing units to storage units within the rack.

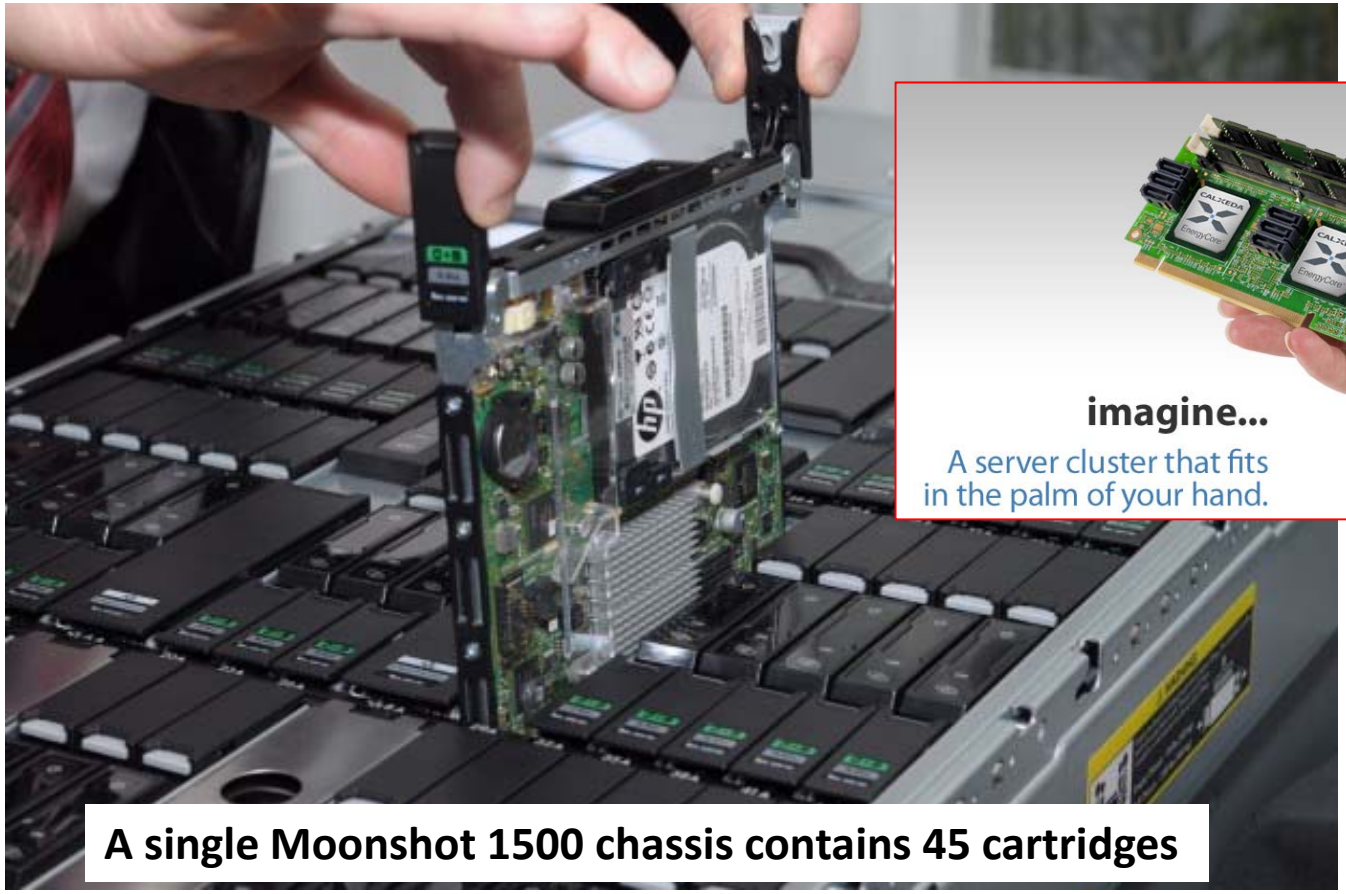


HP Moonshot : “Server Drawer”

HP to Ship 64-Bit ARM-Based Moonshot Server **in 2014**

Compared to traditional servers, up to:

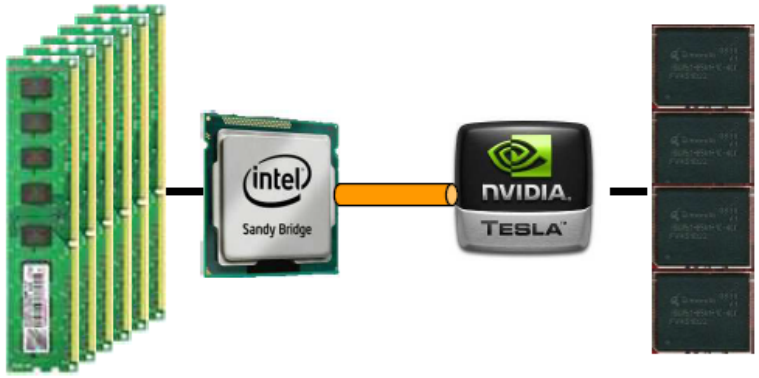
89% less energy * **80%** less space * **77%** less cost ** **97%** less complex *

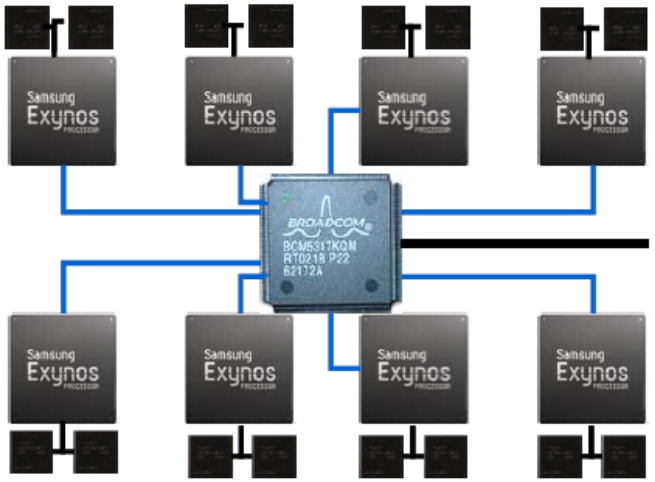


A single Moonshot 1500 chassis contains 45 cartridges

4 Applied Micro's 64-bit X-Gene chips (ARM) on a single board

What is good about it?





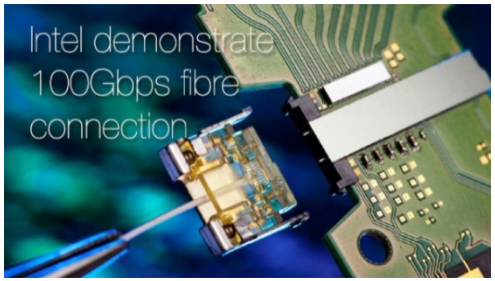
- **Sandy Bridge + Nvidia K20**
  - **1685 GFLOPS**
  - **2 address spaces**
  - 32 GB/s between CPU-GPU
    - 16x PCIe 3.0
  - 68 + 192 GB/s
  - **> $3000**
  - **> 400 Watt**

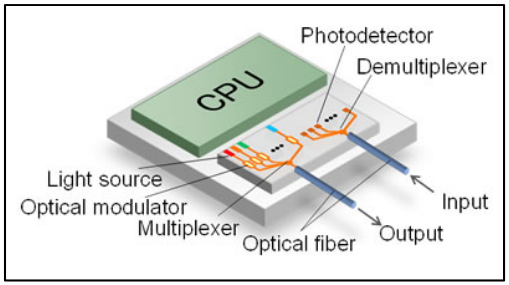
- **8-socket Samsung Exynos 5450**
  - **1600 GFLOPS**
  - **8 address spaces** (ARM+GPU)
  - 12.8 GB/s between CPU-GPU
    - Shared memory
  - 102 GB/s
  - **< $200** **Save x15 $$$$**
  - **< 100 Watt** **Only 25% of the power**

Source: Mont-Blanc Project

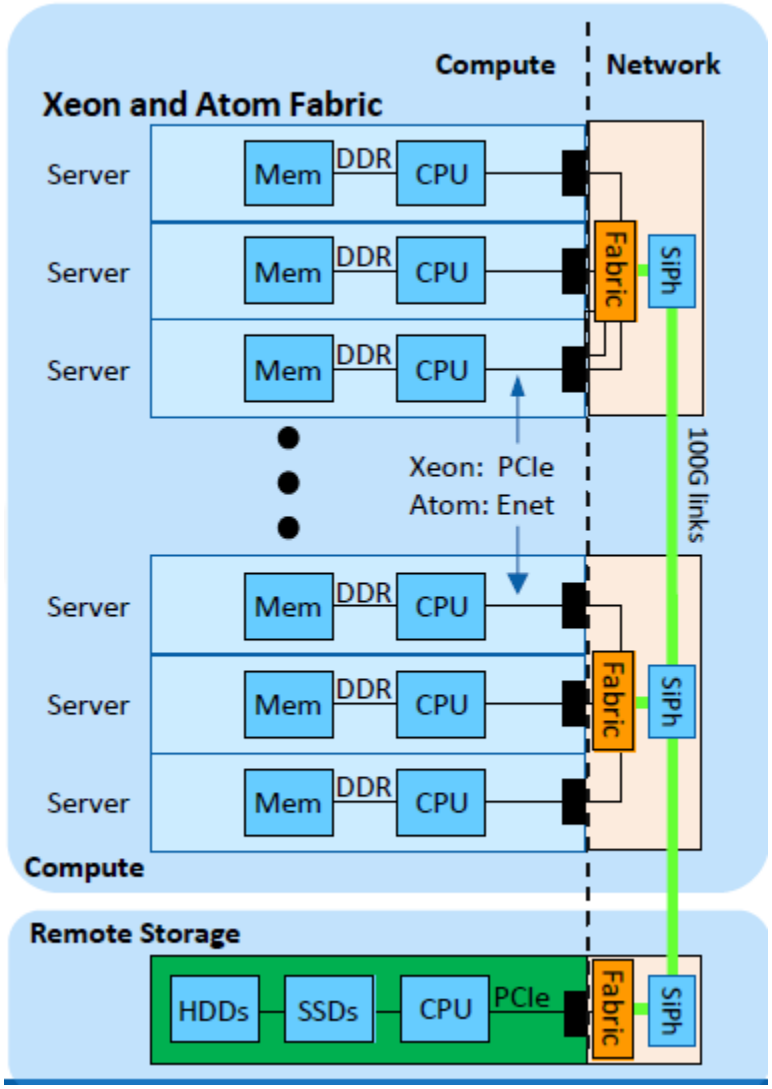
Disaggregated (解耦) rack-scale server architecture

- 资源全面池化
  - **enables independent upgrading** of **compute**, **network** and **storage** **subsystems**
  - **Optics-Based Interconnect**
- **The foundation for the software-defined data center (SDDC)**





IDF2013 @ Beijing

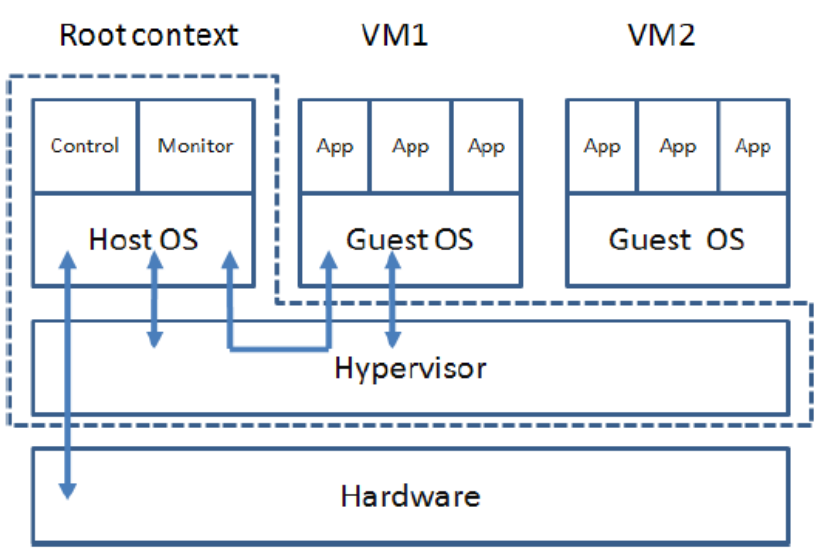


New Design Strategies in OS-1K 解耦操作系统

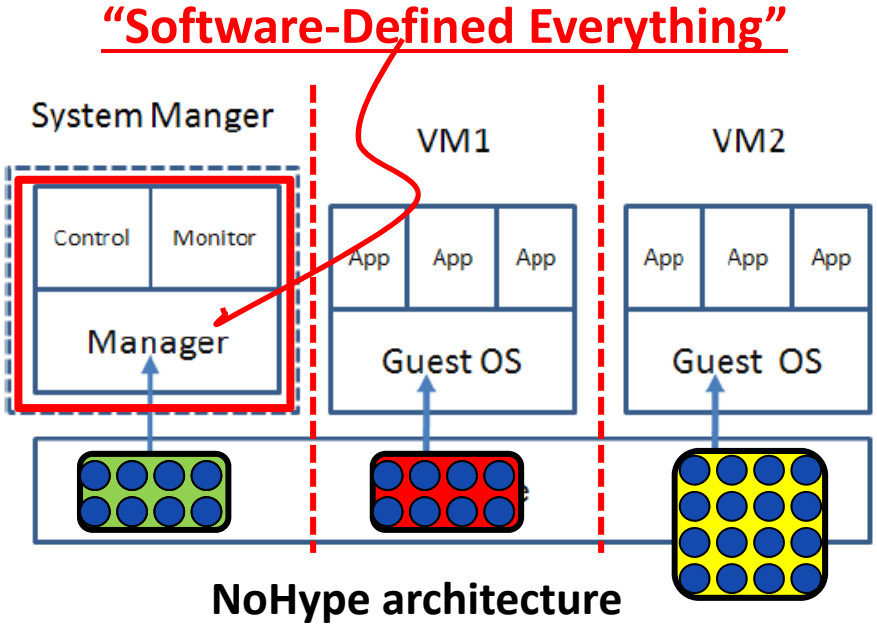
- **Adopt multikernel operating system**
  - **Message passing among kernels to avoid un-necessary chip-to-chip or board-to-board traffic**
- **Space Sharing instead of Time Sharing**
  - **We have "many" cores** 
  - **Stop multitasking → Context switching breaks data locality**
- **"NoHype" Cloud OS (Space-sharing VM)**
  - **Weaker cores (1.0-1.5 GHz) → No hypervisor (Xen, KVM)**
  - **"Resource Slicer" + "Spatial Partitioning".**
- **Compiler or runtime techniques**
  - **to improve data reuse (or increase arithmetic intensity) → temporal locality becomes more critical**

“NoHype” Cloud OS Design

- “NoHype”:
  - Original paper: “Virtualized Cloud Infrastructure without the Virtualization” (ISCA2010) -- Princeton University
    - Additional layer creates security concern
  - Facebook and Google Don’t use virtualization (fully loaded already)
  - “Cores are simpler” in future manycore systems (Intel)

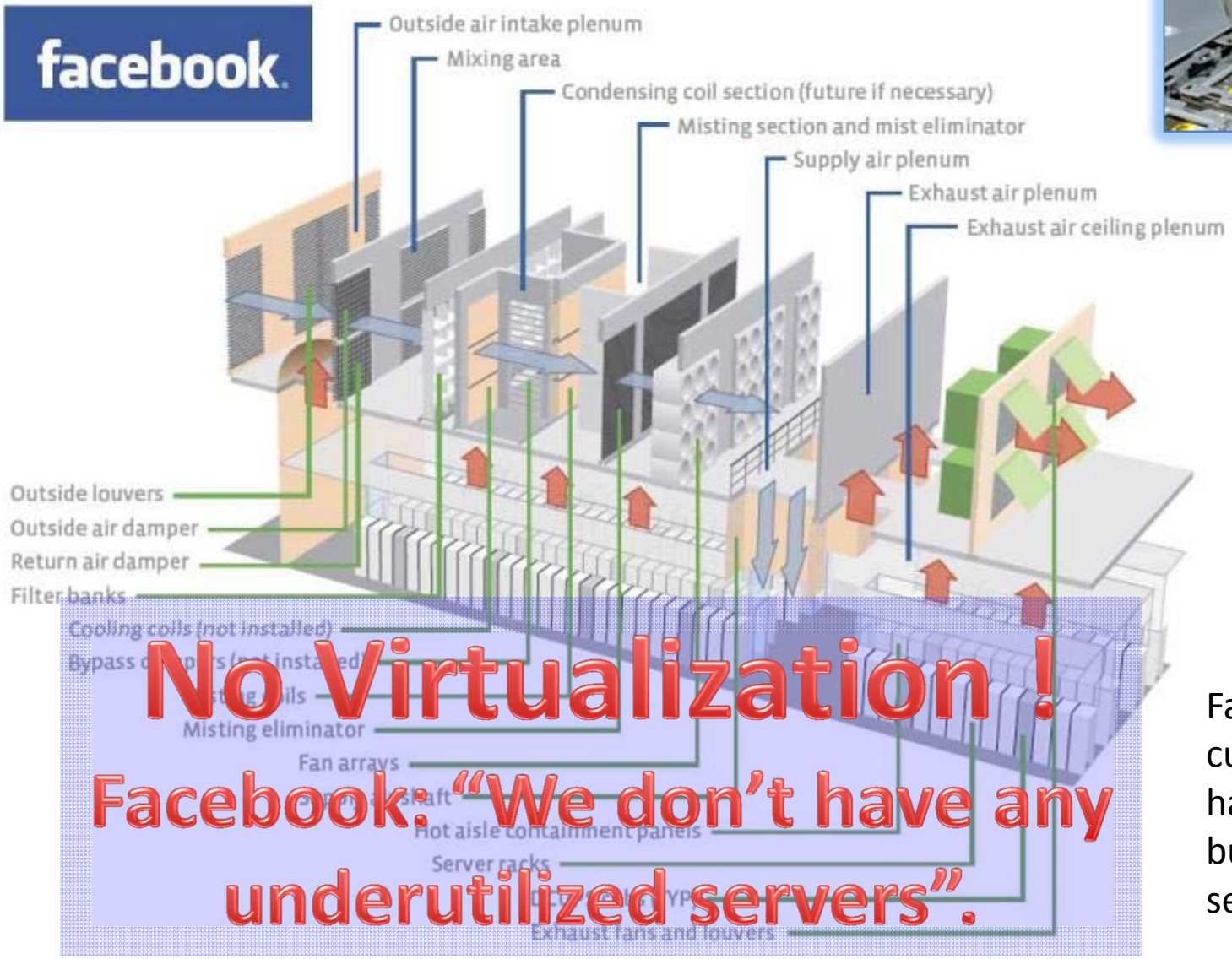


Today’s virtualization layers



NoHype architecture

Open Compute Project Data Center



No Virtualization !

Facebook: "We don't have any underutilized servers".

Facebook went to custom designed hardware instead of buying off the shelf servers from OEMs.

http://www.opencompute.org/

Thanks!

For more information:

C.L. Wang's webpage:
http://www.cs.hku.hk/~clwang/