OpenSolaris and the Direction of Future Operating Systems

James Hughes
Sun Fellow
Solaris Chief Technologist

LISA'08
November 2008
San Diego, CA

Agenda

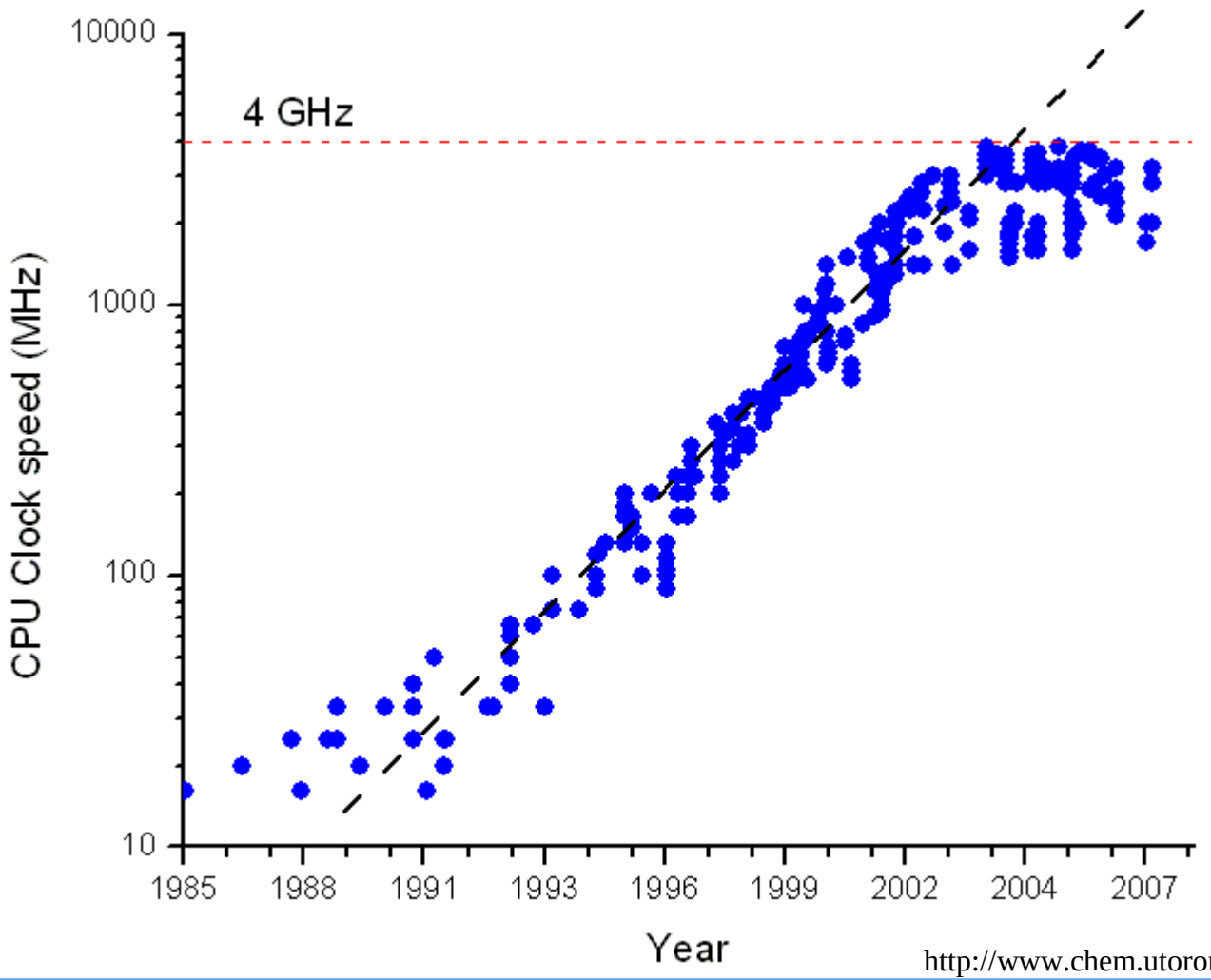
- Operating System Trends
  - Computer / OS architecture trends
  - Why developers matter
  - Direction of the OS
- Solaris and OpenSolaris
- Security features
  - eZFS, Xlofi
  - Key Management
  - Containment

What is an Operating System?

The Future of Operating Systems

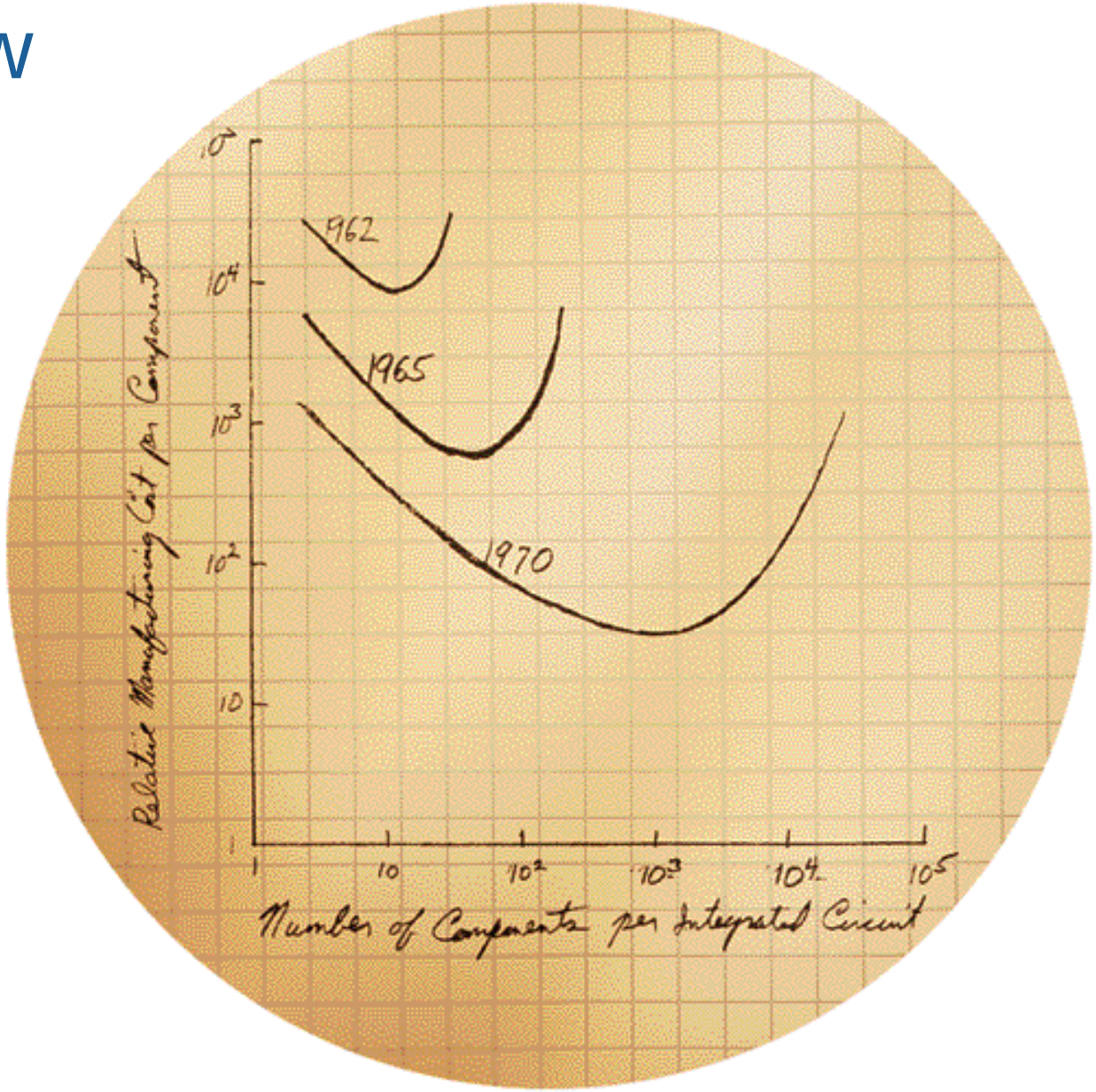
- Hardware Trends
  - Moore's Law
- Operating Systems Trends
  - Large Scale
  - NUMA
- Programming Trends
  - OpenMP
  - Fortress
  - MapReduce (Phoenix, Hadoop)

Clock Speeds



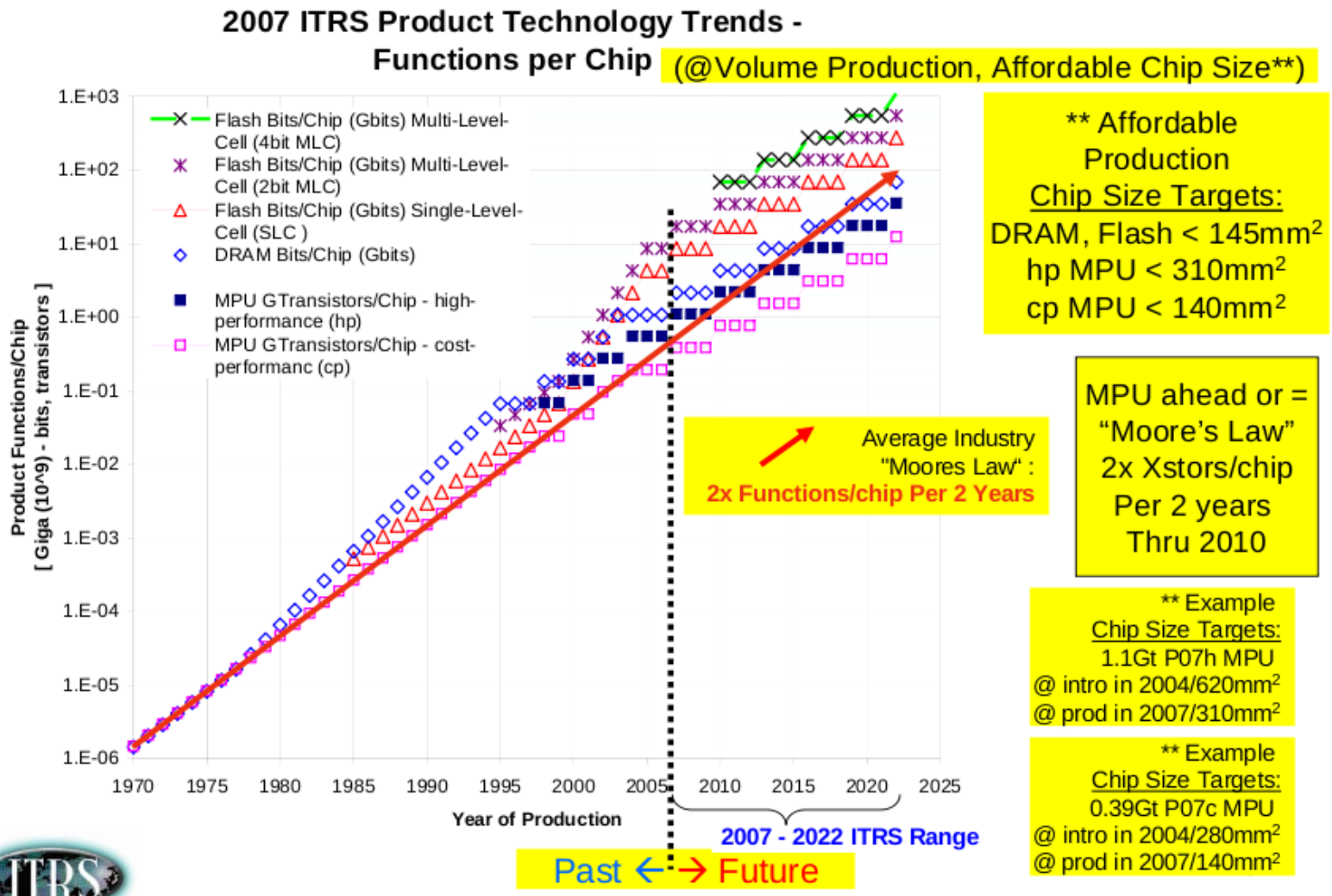
http://www.chem.utoronto.ca/~nlipkowi/

Moore's Law

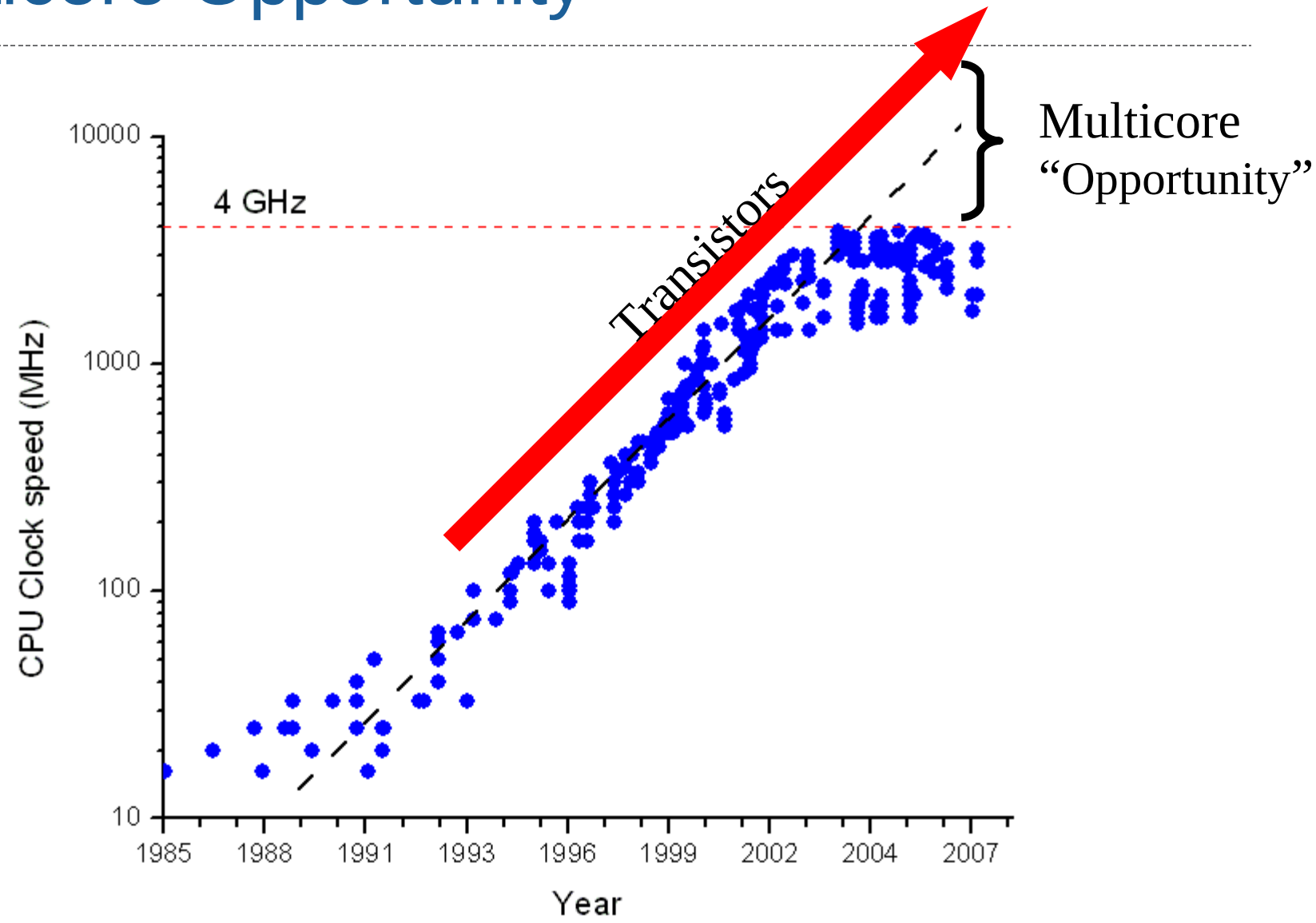


http://www.intel.com/museum/archives/history_docs/mooreslaw.htm

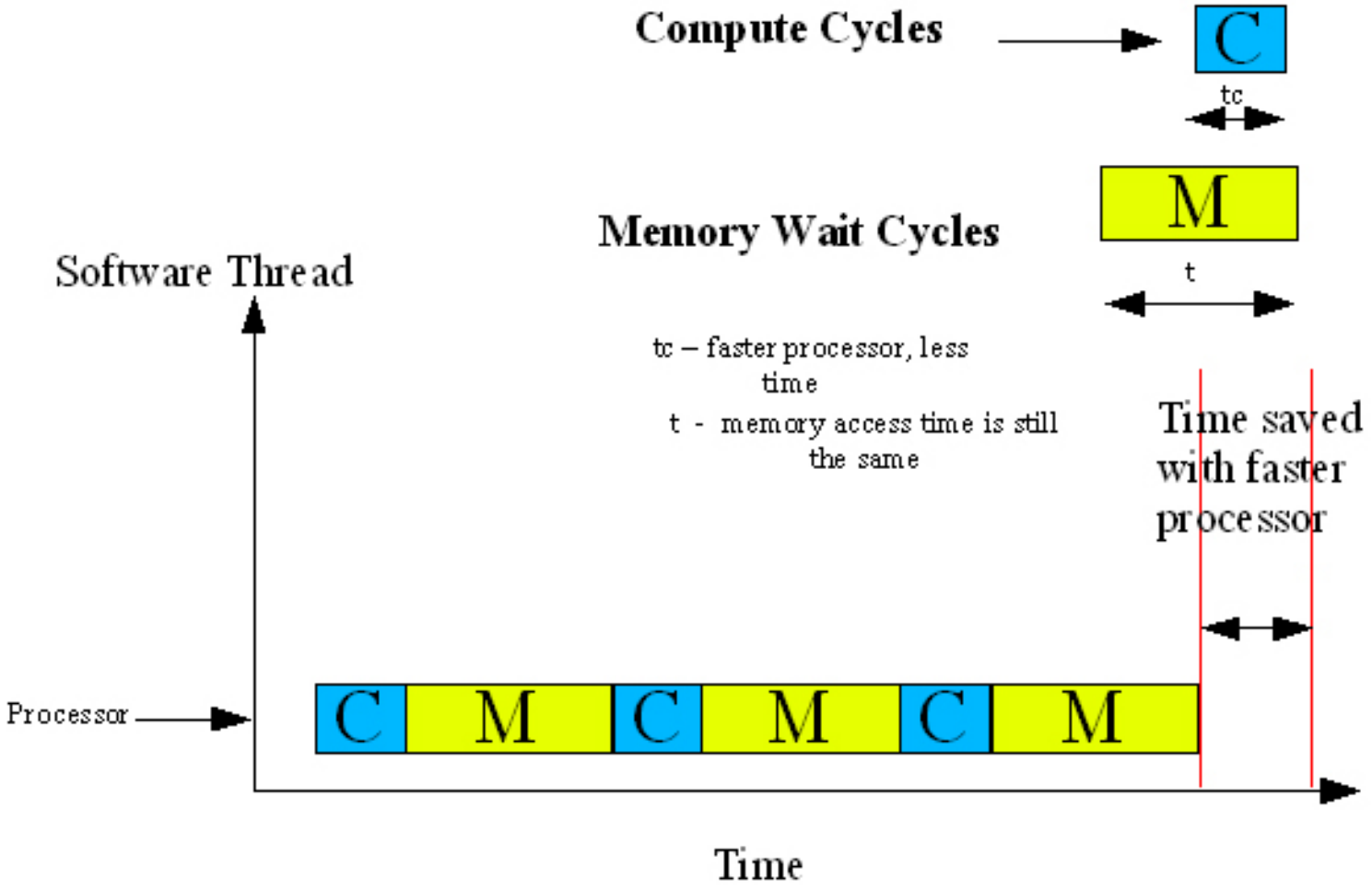
Chip Size Trends – 2007 ITRS Functions/Chip Model



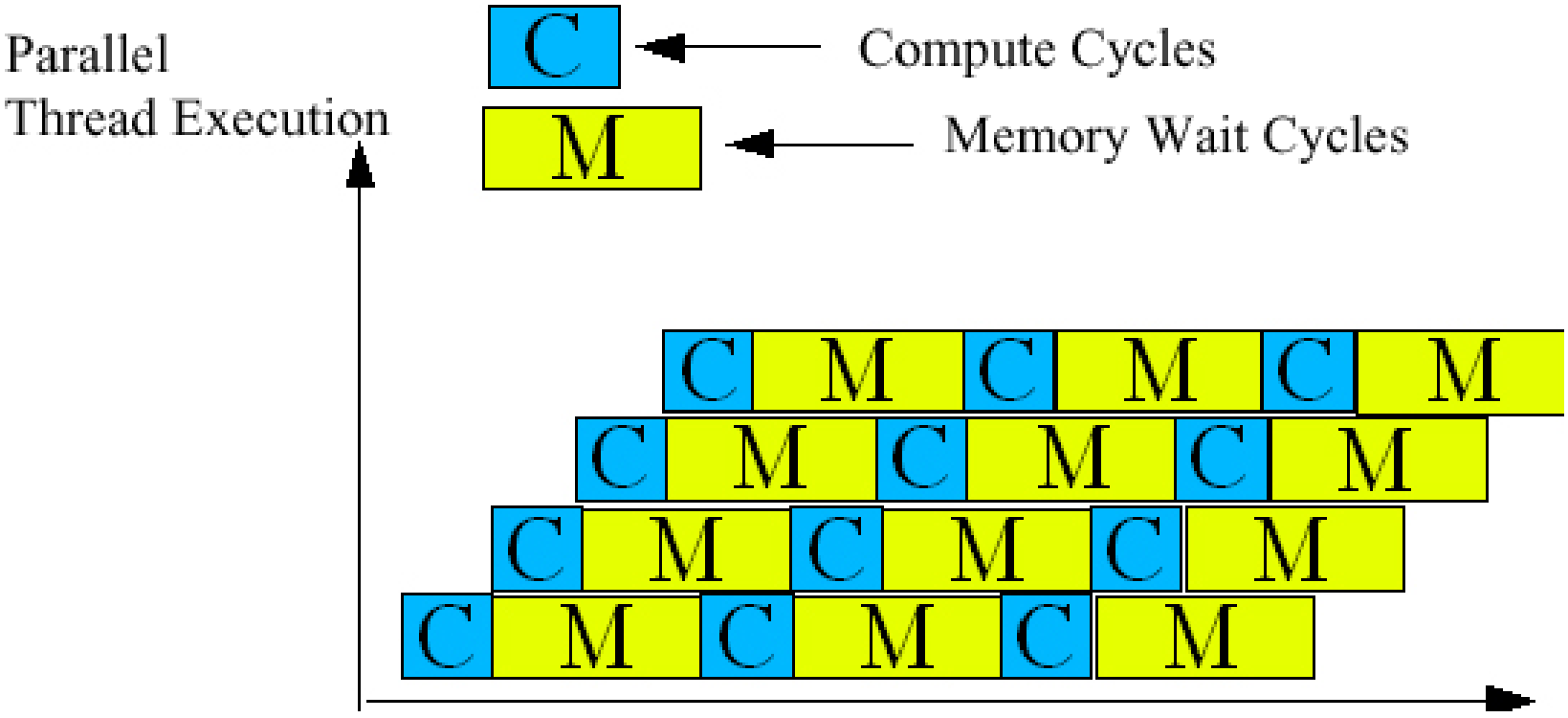
Multicore Opportunity



Memory Latency

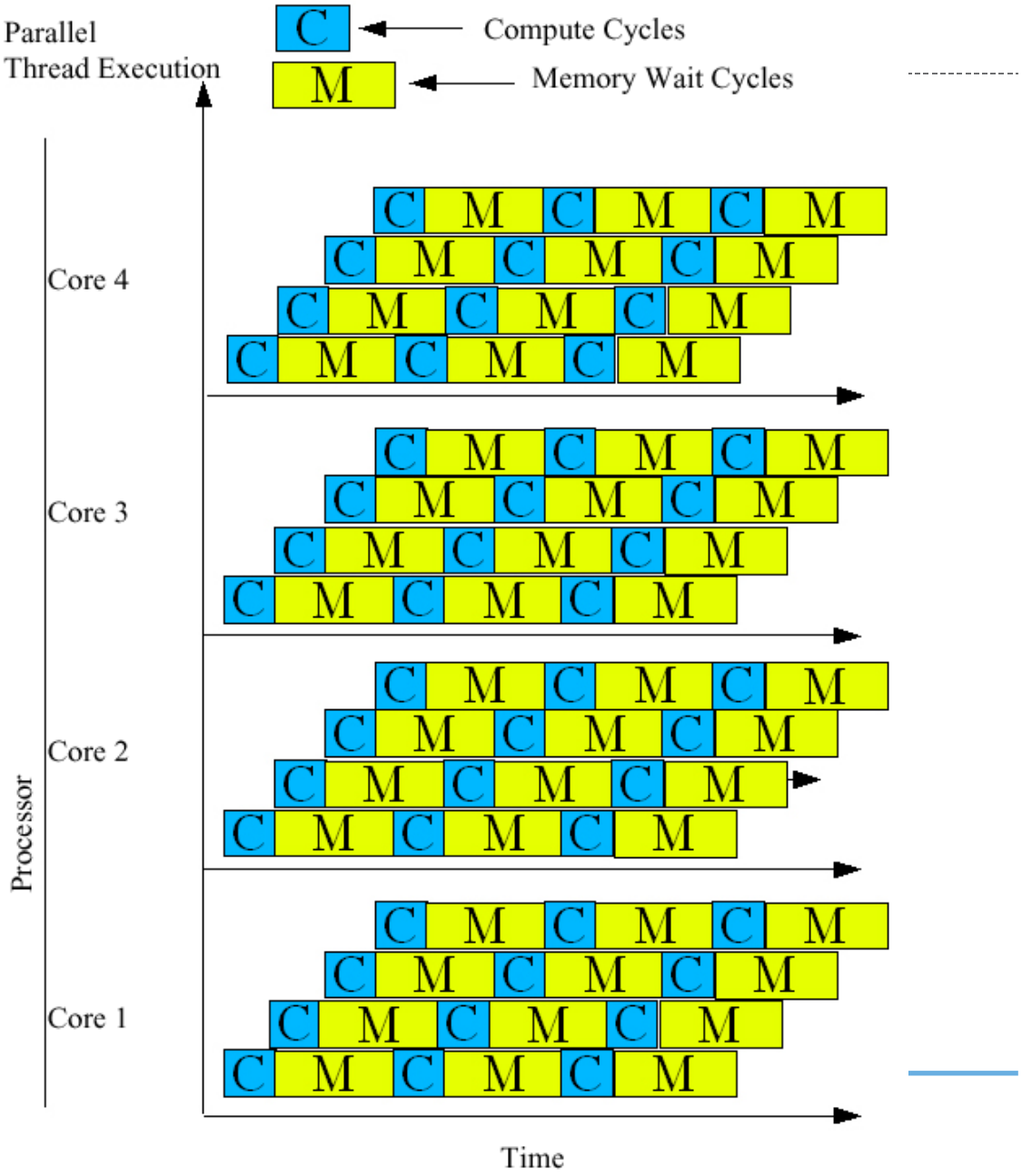


Multi threading



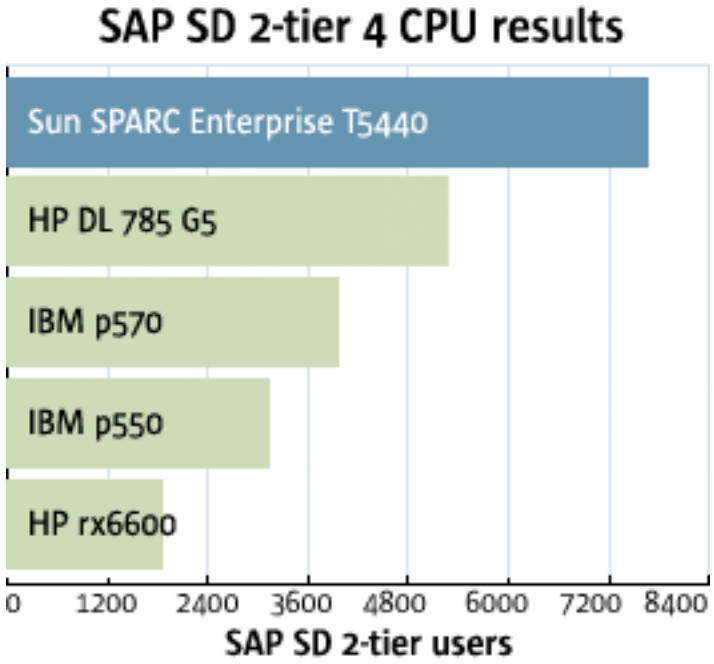
CMT

- Multicore Opportunity
- Leverages Multicore and Multi threading



Batoka (aka T5440)

- 256 hardware threads
- 0.5 TB of RAM
- 64 integer units
- 32 floating point units
- 32 crypto accelerators



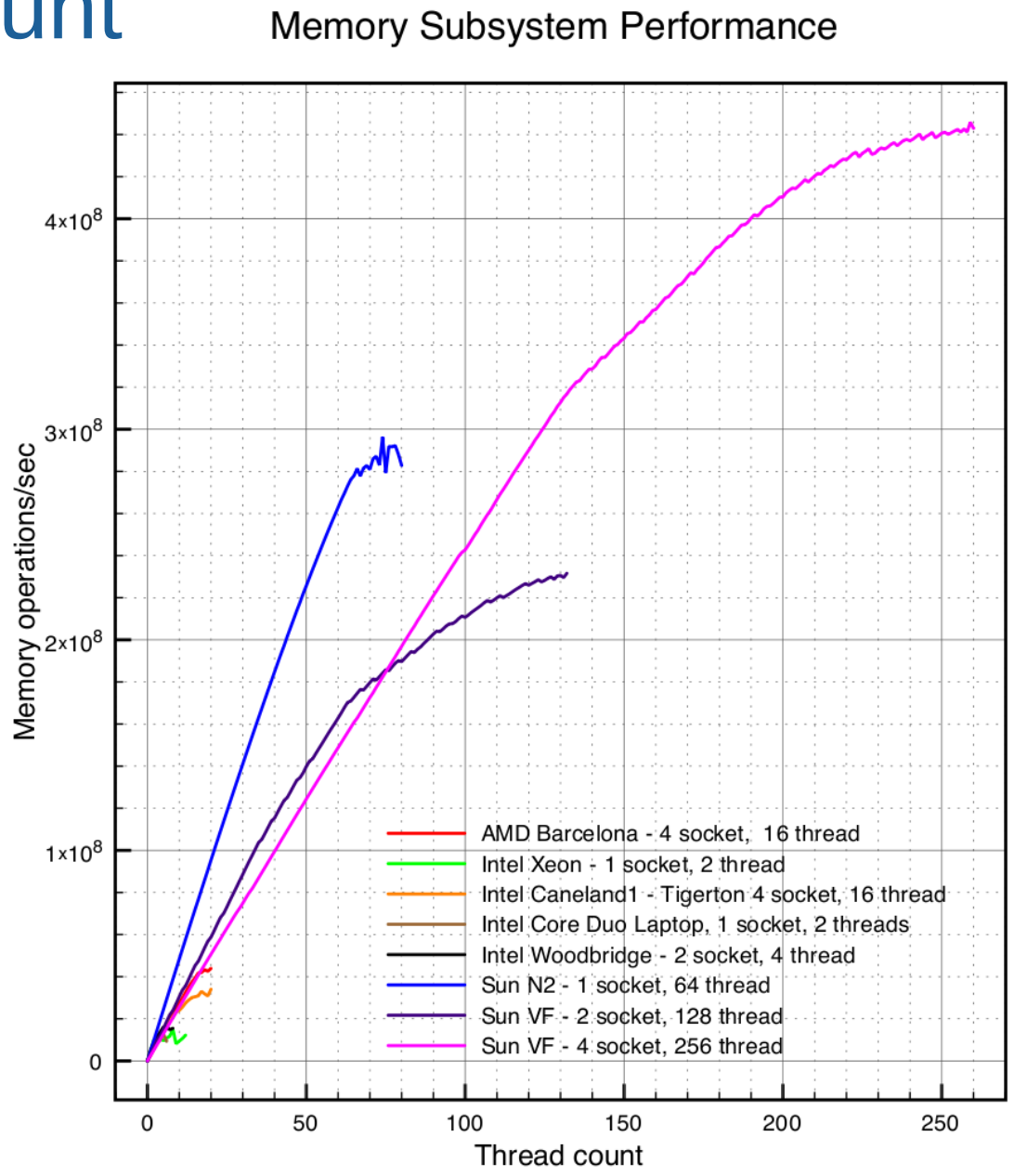
4u

Searching for Goldilocks Applications

- Not to large
  - Blows iCache
- Not to small
  - Uninteresting
- Just right
  - Phenomenal Performance
- Pheonix Map Reduce
  - Free, SPARC, CMT

High Thread Count

Memory Ops/sec
vs
Thread count



The future is going to be high thread counts

The winners will solve the problems using parallel methods

Programs will parallelize at runtime

Languages will hide parallelism from the programmer

Parallel apps are becoming real

- Hot
  - Map Reduce
    - Hadoop
    - Phoenix
  - Fortress
- Not
  - MPI, OpenMP

Operating Systems provides the...

...glue between the application and the hardware

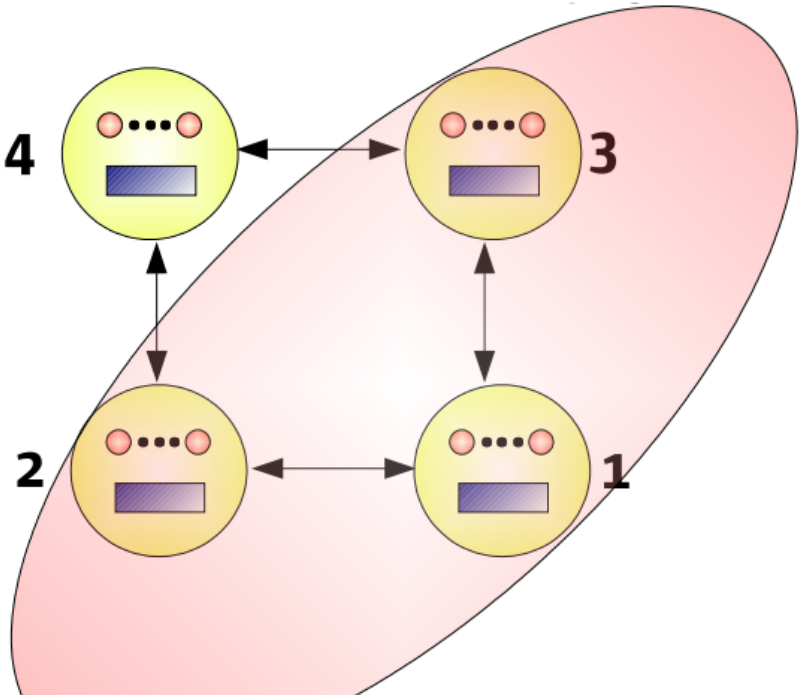
...application tools and libraries needed to get its job done

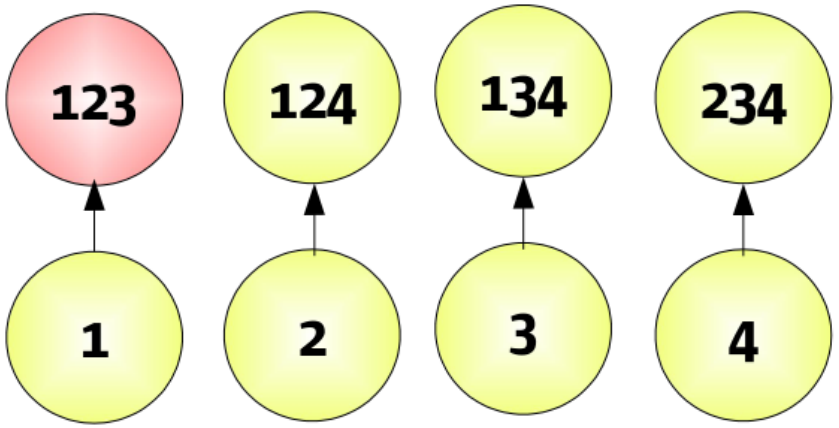
...programmer's productivity for development and debugging

- Operating Systems have to reduce complexity
  - While enabling efficiency

Managing NUMA

- Helping the programmer be efficient
  - Transparent to the programmer; Simple
- Managing locality of memory
- OpenSolaris “Memory Placement Option”





ZFS

- Revolutionary file system
- Data Integrity
- Encryption (soon)
- Simple
- Allows “Mulligans”

Dtrace

- Allows applications debugging on production code
- Allows logic to be executed in the traps
- Takes a complicated debugging problem and makes it simple.

Operating System Future

- Efficiency -and- Capability
- Enable applications that require large memory footprint and high thread counts
- Allowing applications to be bigger, faster
- Solve numerically hard problems
  - Simulations
  - Financial models
  - Physics simulations (aka games)

Languages for Parallelism

- Explicit Parallelism for Clusters
  - OpenMPI
  - Map Reduce
    - Hadoop
- Explicit Parallelism for SMP
  - OpenMP
  - Cilk
  - Fortress
  - Map Reduce
    - Phoenix

Cilk

- Cilk is an algorithmic multithreaded language
- Cilk is algorithmic
  - guarantees efficient and predictable performance
- Runs on OpenSolaris

```
cilk int fib (int n)
{
    if (n < 2) return n;
    else
    {
        int x, y;
        x = spawn fib (n-1);
        y = spawn fib (n-2);
        sync;
        return (x+y);
    }
}
```

Fortress

- A new programming language designed for high-performance computing (HPC) with high programmability.
  - Implicit parallelism
  - Transactions
  - Flexible, space-aware, mathematical syntax
  - Static type-checking (but with type inference)
  - Definition of large parts of the language in its own libraries

Describe algorithms in math terms

$z = 0$
$r = x$
$\rho = r^T r$
$p = r$
DO $i = 1,25$
$q = A\,p$
$\alpha = \rho / (p^T q)$
$z = z + \alpha\,p$
$\rho_0 = \rho$
$r = r - \alpha\,q$
$\rho = r^T r$
$\beta = \rho / \rho_0$
$p = r + \beta\,p$
ENDDO
compute residual norm explicitly: $\|r\| = \|x - A\,z\|$

$z : \mathrm{Vec} = 0$
$r : \mathrm{Vec} = x$
$p : \mathrm{Vec} = r$
$\rho : \mathrm{Elt} = r^{\mathrm{T}} r$
for $j \leftarrow \mathbf{seq}(1 : cgit_{\max})$ **do**
$q = A\,p$
$\alpha = \dfrac{\rho}{p^{\mathrm{T}} q}$
$z := z + \alpha\,p$
$r := r - \alpha\,q$
$\rho_0 = \rho$
$\rho := r^{\mathrm{T}} r$
$\beta = \dfrac{\rho}{\rho_0}$
$p := r + \beta\,p$
end
$(z\,,\ \|x - A\,z\|)$

Map Reduce

1) Map input to (key, value)
2) Sort by key
3) Reduce (key, value) to the solution

- Steps 1 and 3 describe functions that are independent to scaling
- Parallelism is differed to run time
  - Possibly without the programmers knowledge

Map reduce Implementations

- Hadoop
  - Java based
  - Scales wide
- Pheonix (Stanford)
  - C based
  - Scales to high threads
  - Large memory
  - Best when problem fits in memory
- Both available for OpenSolaris

“…but my application can't scale”

If you don't parallelize your applications, your competition will

Information Technology can be a competitive advantage

Solaris 10 and OpenSolaris

- Enterprise quality and support
  - New Hardware
  - Compatible change
- OpenSolaris
  - New capabilities
  - New management strategies
  - Community driven
  - Developer support

Solaris 10 update 6

- ZFS Root / boot
  - New SPARC boot loader
- ZFS Deligated administration
- Default IP route per zone
- SHA 256/512
- 256 hardware threads on x86
- Performance improvements
- Updates of S10 will continue

OpenSolaris 2008.05

- Leading indicator of Solaris
  - ZFS root
  - New packaging
  - New patching
  - Familiar userland
- Support for developers
- Updates every 2 weeks
  - “New” every 6 months

OpenSolaris 2008.11

- New features
  - TimeSlider
- New and refreshed OpenSource
  - sudo and others
- New repositories
  - Redistributable repository
  - Community repository
  - Closed repository

Security matters to the developer

- Direction for OpenSolaris
  - Encrypted Storage
    - Tape (now)
    - eZFS
    - xlofi
  - Key Management
  - High Assurance containment
    - Windows in a Solaris TX labeled zone

Storage contains personal information

- California law about data breaches
  - Many examples
- Laptops being lost
- Thumb drives
- Cell phones
- Storage contains
  - All communications
  - Work in progress

Today

- When the user is not logged in,
  - the administrator ***can*** see the data
- With ZFS and Enterprise RAID,
  - overwriting a file does not erase the data
- Data on RAID is clear on single disk
  - In **`m`** of **`n`**, **`1/n`**th of the data on each disk

eZFS

- When the user is not logged in,
  - the administrator ***can not*** see the data
- Zeroing the key erase the data
  - Permanently
- Data is protected regardless of strategy
  - RAID, mirror, etc.

Backup should be under separately managed key so that users are not vulnerable to key loss

Future

- eZFS
  - Encrypted Boot
- Xlofi
  - Turns file (partition, zVol, etc.) into secure disk
- Batoka – 12GBytes/s of AES

"All Storage leaves the datacenter one way or another, sooner or later"

Key Management

- Requirements are simple
  - “Don't lose the keys”
  - “Don't give the keys to the wrong people”
- OOB key requirement
- Many organizations working on this
  - Companies, Standards, etc.

How do you manage your keys?


CARLSEN
A
SLAVE

Sample Customer

- 100,000 individual keys
  - Today!
- Auditors having a fit
  - Used to copy information between servers for batch processing
- Do you know a customer that has this problem?
- Solution is not high tech
  - Capture, categorize, manage, whole lifetime

Encrypted Storage vs HW Trends

- Measured AES, 100MB/s, on Laptop
  - AMD, Intel and Sun will have acceleration
  - Batoka, 12GBytes/s
- Single disk performance 40MB/s (not Flash)
- First access has latency
  - Subsequent access access in RAM buffer
- This level of performance is “free”
  - In the OS is “free”
- “Security is an expectation, not a market”

Long Term Prediction of Adoption

- Computers are fast enough
- OS vendors will add for free
  - Yes, there are country issues
- At least password protected
- There is no reason not to encrypt

In the future, not encrypting your storage will be like using telnet instead of ssh

Solaris TX

- Military grade Sandbox of
  - Applications (aka Zones)
  - Virtual Machines (VB in TX)
  - Extends to throughout the datacenter over the networks
- MLS Applicable to more that Governments
  - Servers that handle high value transactions

Conclusion

- Operating Systems
  - Computer / OS architecture trends
  - Why developers matter
  - Direction of the OS
- Solaris and OpenSolaris
- Security features
  - eZFS, Xlofi
  - Key Management
  - Containing Windows

Thank you

James.Hughes@sun.com