Redshift: The Explosion of Massive Scale Systems

Dr. Greg Papadopoulos
Executive Vice President, R&D
Chief Technology Officer
These slides contain forward-looking statements regarding the future results and performance of Sun Microsystems, Inc., including statements regarding future industry trends and growth, demand for HPC and future trends in computing. These forward-looking statements involve risks and uncertainties, and actual results could differ materially from those contained in these forward-looking statements. Factors that could cause actual results to differ materially from those contained in these forward-looking statements include: risks associated with developing, designing, manufacturing and distributing new products; lack of success in technological advancements; pricing pressures; lack of customer acceptance of new products; the possibility of errors or defects in new products; competition; adverse business conditions; failure to retain key employees; the cancellation or delay of projects; our reliance on single-source suppliers; risks associated with our ability to purchase a sufficient amount of components to meet demand; inventory risks; risks associated with our international customers and operations; delays in product development or customer acceptance and implementation of new products and technologies; our dependence on significant customers and specific industries; and our dependence on channel partners. Please also refer to Sun's periodic reports that are filed with the Securities and Exchange Commission, including Sun's annual report on Form 10-K for the fiscal year ended June 30, 2006 and its quarterly report on Form 10-Q for the fiscal quarter ended October 1, 2006. Sun assumes no obligation to, and currently does not intend to, update these forward-looking statements.
25 Yrs: 100,000 to 1M Times More MIPS/$

Hermann Brunner, Max-Planck-Institut für Extraterrestrische Physik, Germany
Where is the Demand?

Processor Performance

Sun 1
1 MIPS
$40K (2005$)

T1000
10,000 MIPS
$5K (2005$)

$17T

World GDP

$36T
Core per Enterprise Demand

Moore’s Law

Core Enterprise Demand

1995  2000  2005  2010
“Last Mile” Bandwidth
ΣBW – Filling the Pipe

ΣBW

In = Out
$\Sigma BW$ – Filling the Pipe

\[ \text{In} = \text{Out} \]

$\# \text{Servers} \times \text{BW/Server} = \# \text{Devices} \times \text{BW/Device}$
ΣBW – Filling the Pipe

- **ΣBW**: The throughput of the data pipe
- **Moore’s Law**: The rate of increase in computing power
- **Core**: The rate of increase in the number of cores
HPC – Insatiable Demand
HPC – Insatiable Demand

- Moore’s Law
- Core
- ΣBW
- HPC
*-Prise

Internet

salesforce.com

Workday

Network.com

ebay

amazon.com

Gmail

webex

Windows Live
*Prise

Moore’s Law

ΣBW

HPC

*Prise

Moore’s Law

CRM

Core
Redshift – a Move to Massive Scale

- **Moore’s Law**
- **Core**
- **ΣBW**
- **HPC**
- **Prise**

The graph shows the following trends over the years 1995 to 2010:

- Moore’s Law: A steady increase in processing power.
- Core: A consistent growth in performance.
- **ΣBW**: Significant growth in bandwidth.
- **HPC**: Rapid increase in high-performance computing capabilities.
- **Prise**: Steep rise, indicating a rapid advancement in technology.

The data highlights the exponential growth in various technological parameters, aligning with the concept of Redshift – a move to massive scale in computing and data processing.
Diverging Concerns

- Moore’s Law
- HPC
- *Prise
- Core
- Consolidation Cost Services

Graph showing exponential growth from 1995 to 2010.
Diverging Concerns

- **Scale Efficiency Products**
- **Moore’s Law**
- **Core**
- **Consolidation Cost Services**
“The Google Question”

If all you cared about was massive scale, what would you build and how would you operate it?
An Answer

With “Brutal Efficiency”

• Hardened metrics (utilization, power, security)
• Predictability of service level
• Idea-to-deploy time and productivity
An Answer

With “Brutal Efficiency”

- Hardened metrics (utilization, power, security)
- Predictability of service level
- Idea-to-deploy time and productivity

Deep, At-scale Engineering of Systems
How Deep?

Efficiency and predictability at massive scale are as mission-critical to Redshift as RAS has been to the core enterprise.
Don’t Confuse

Commoditization of Computers
Don’t Confuse

Commoditization of Computers

With

Commoditization of Computing
System Includes...

- Base HW Plant (Servers, Storage, and Switches)
- O/S Instances
- Core Services and Platforms
- Network Service
System Includes...

Massively Parallel System

Network Service

Core Services and Platforms

O/S Instances

Base HW Plant
(Servers, Storage and Switches)
**SMPs Are Back (With a Vengeance!)**

64 Threads in 1997

**E10K**

<table>
<thead>
<tr>
<th>Full Rack</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>9620 Watts</td>
<td>Power (Systems at peak utilization)</td>
</tr>
<tr>
<td>1,800 lbs.</td>
<td>Weight</td>
</tr>
<tr>
<td>~150k tpm</td>
<td>Performance</td>
</tr>
</tbody>
</table>
## SMPs Are Back (With a Vengeance!)

<table>
<thead>
<tr>
<th></th>
<th>Size</th>
<th>Power</th>
<th>Weight</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Full Rack</td>
<td></td>
<td>9620 Watts</td>
<td>1,800 lbs.</td>
</tr>
</tbody>
</table>

### 1997

- **E10K**
  - 64 Threads

### 2007

- **Niagara 2**
  - 64 Threads
  - Weight: 1,800 lbs.
  - Power: 9620 Watts
  - Performance: ~150k tpm
SMPs Are Back (With a Vengeance!)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>Full Rack</td>
<td>1U</td>
</tr>
<tr>
<td>Power</td>
<td>9620 Watts</td>
<td>410 Watts</td>
</tr>
<tr>
<td>Performance</td>
<td>~150k tpm</td>
<td>~300k tpm</td>
</tr>
<tr>
<td>Weight</td>
<td>1,800 lbs.</td>
<td>45 lbs.</td>
</tr>
</tbody>
</table>

- 30x reduction in size
- 24x reduction in power (Systems at peak utilization)
- 40x reduction in weight
- 2x increase in performance
SMPs Are Back (With a Vengeance!)

- 64 Threads in 1997
  - E10K

- 64 Threads in 2007
  - Niagara 2
Neptune and Crossbow

Crossbow
- Solaris virtual networking stacks for high speed networks, based on protocol, service, or container
- Does the “careful” unleashing of 10G

Neptune
- A shared, virtualized, non-blocking, multi-homed 10 Gigabit Ethernet PCI-e network interface
- Gets the most out of your system and your I/O bus
  > Processors get faster faster than I/O buses get faster – PCI-e is (or soon will be) your bottleneck
Neptune and Crossbow

Solaris Zones Processes

Solaris Crossbow and PEF

MultiCore Microprocessor

Virtualized Network Streams

PCle

Neptune

1000’s Logical Connections

4x10GE

1000’s Logical Connections
Java RTS + Solaris = Open Real Time Operating System
Java RTS + Solaris: 1000 Times More Predictable

<table>
<thead>
<tr>
<th></th>
<th>Maximum Latency</th>
<th>Jitter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solaris</td>
<td>20 microsec</td>
<td>10 microsec</td>
</tr>
<tr>
<td>Linux</td>
<td>10 millisec</td>
<td>5 millisec</td>
</tr>
<tr>
<td>Windows</td>
<td>12 millisec</td>
<td>7 millisec</td>
</tr>
</tbody>
</table>
Real-Time Application Server Results
Comparison of Conventional Java SE to Java RTS

Distribution of Request/Reply Round-trip Times

**Java RTS Results:** ALL priority transactions complete in 11 milliseconds or less

**Standard Java SE Results:** Priority transactions range from 11 ms to 3.5 seconds
Real-Time Java on Solaris: Inverted Pendulum

Greg Bollella
Distinguished Engineer, Sun Microsystems
NetBeans IDE Support

Download the Java RTS Plugin, and

• Cross-develop on the host
• Deploy over the network
• Execute on the target
• ...from the NetBeans IDE
Co-Design

Computer + Storage + Network + Power + Cooling + Software = BLACKBOX
But Wait, There’s More

- The next wave will include leading-edge enterprises
- Critical to this wave will be the bridge between core IT and redshifting services:
  - Identity and security
  - Procedural languages and scripting
  - SOA and Web 2.0
  - New clients
The Bridge

- HPC
- *Prise
- Identity
- RSWS
- New Clients

Graph showing growth from 1995 to 2010.
Enterprise Bridge
Putting it all Together: The Binary Distribution

“The Distro”

Base HW Plant (Servers, Storage and Switches)

O/S Instances

Core Services and Platforms

Network Service
Source vs. Binary

Platform Developers

Platform
(Open Source)

opensolaris
open Java
open ...
Source vs. Binary

Platform Developers

Platform (Open Source)

Distribution

opensolaris
open Java
open ...

Java
solaris ...

Binary Distribution

1.1 → 1.2 → 1.2.1
Source vs. Binary

Platform Developers

Platform (Open Source)

Application Developers

Applications

open solaris
open Java
open ...

Distribution

Binary Distribution

1.1

1.2

1.2.1

Application Binaries
A Likely Scenario

Big Computers
100ks – 1Ms Nodes
Globally Distributed

Devices
A Likely Scenario

Big Computers
100Ks – 1Ms Nodes
Globally Distributed

Distribution Networks

Devices
A Likely Scenario

Big Computers
100Ks – 1Ms Nodes
Globally Distributed

Distribution Networks

Startups

Devices
Thank You.
The Network is the Computer.™

Dr. Greg Papadopoulos
Executive Vice President, R&D
Chief Technology Officer