Oracle Solaris - The Best Platform to run your Oracle Applications

David Brean
Oracle Solaris Core Technology
Safe Harbor Statement

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle’s products remains at the sole discretion of Oracle.
Solaris 11. The First Cloud OS.

Next Generation of UNIX

• #1 UNIX
  – Industry leading availability, security and performance

• Built for Cloud Infrastructures
  – Breakthrough architecture to deploy and secure and manage enterprise clouds

• Engineered for Oracle
  – Optimized for Oracle software and hardware
Oracle and Oracle Solaris

Accelerate Access to Hardware Innovation

Co-Engineered for Innovation

Creating unique value for Engineered Systems

Oracle and Oracle Solaris

Accelerate Access to Hardware Innovation

Co-Engineered for Innovation

Creating unique value for Engineered Systems

Oracle and Oracle Solaris
Oracle Solaris 11
Unique Value for Oracle Engineered Systems

**Exadata**
- Optimized Shared Memory
- Integrated auditing
- InfiniBand optimizations
- 2.5x faster reboots

**Supercluster**
- Zero overhead virtualization
- Subsecond application failover
- 5x faster encryption
- Deduplication for 10x storage savings
- 4x faster upgrades

**Exalogic**
- 6x faster cloud deployments
- Integrated load balancer, router, firewall
- DTrace for Solaris and Java for unmatched observation
- Assignable network bandwidth per application
- Unique InfiniBand and Zones integration
Limited choices is a good thing!

Engineered systems do the important things VERY well with uniformity

- System configuration locked down
  - Hardened, tested together, guaranteed to work well
  - Software can be optimized to specific known configuration

- Standardized datacenter footprint & building block
  - Standardized power, cooling and network integration

- Interoperability and expandability
  - Exalogic, Exadata, SPARC SuperCluster and other Engineered Systems are designed to work together
  - Storage and networking expansion blueprints for more flexibility
Full Stack SW Development

- Co-engineered lifestyle
  - Enhanced performance, unique features, built-in robustness, observability, security

- Tested together
  - 150,000 machine hours every week, 10M+ machine hours annually

- Coordinated patching
  - Easy, risk free updates

- Supported together
  - Faster problem resolution
InfiniBand Fabric

Infiniband technology is the heart of Engineered Systems

- Infiniband Quad Data Rate (QDR) runs at 40Gbps
  - Full line-rate non-blocking from all components
- Extremely low latency – 10x better than 10GbE
- Redundancy on all paths – auto failover, non-service affecting
- Connects compute, storage and interconnects all Engineered Systems from Oracle – Exalogic, Exadata, SPARC SuperCluster
Exalogic I/O and Network Design

Exalogic X2-2

- Ethernet Gateway Switches
- Spine Switch
- Compute Nodes
- Storage

Management Switch

Data Center Mgmt Network (GbE)
- GbE

Data Center Service Network (10GbE)
- 10GbE

Exabus
- InfiniBand I/O Backplane
- EoIB/SDP/IPoIB/SDP/IPoIB

SDP/IPoIB

Exadata Exalogic SPARC SuperCluster

Management Network (GbE)

Standard Oracle Database

Oracle Database

Exbus

Exabus (InfiniBand I/O Backplane)
The InfiniBand Stack
IP over InfiniBand (IPoIB)

- IETF standard compliant
  - 4391 – IP transmission over IB
  - 4755 - IP transmission over the RC and UC modes of IBA to support large MTUs
  - 4390 – DHCP over IB
- 20 byte link layer address
- Maps IP subnet to IB partition
- Enables IP addressing for other IB ULPs

- TCP checksum offload and LSO, interrupt moderation supported
- HA Support: IPMP
- Observability
  - Wireshark for IB
The InfiniBand Stack

Encapsulation

• Ethernet over IB (EoIB)
  – Protocol used with the Sun Gateway (NM2-GW) for bridging IB and Ethernet networks
    • Protocol defines the procedure for tunneling both unicast and multicast Ethernet packets over IB using the Unreliable Datagram (UD) transport.
    • The Ethernet packet includes the standard header, VLAN tags, if appropriate, and the payload.
      – Protocol defines control messages used by Gateway management software to associate IB connected servers with Ethernet ports on Gateway
  – EoIB protocol server-side endpoint in OS is implemented as standard network driver (NIC)
    • EoIB NICs behave like regular Ethernet NICs
    • Networking stack works seamlessly over EoIB NICs
The InfiniBand Stack
Sockets Directs Protocol (SDP)

• Access to IB via Socket API
• IBTA standard wire protocol
• SOCK_STREAM transport
• Leverages InfiniBand Capabilities
  – Transport Offload – Reliable Connection
  – Zero Copy – Using RDMA
  – Kernel Bypass
  – Delivers low latency
• netstat(1m) extensions
The InfiniBand Stack

Reliable Datagram Sockets (RDS)

- Protocol used exclusively by the Oracle Database
- Provides in order, non-duplicating, highly available, low overhead, reliable delivery of datagrams between hundreds of thousands of non-connected endpoints
  - Datagram protocol like UDP but reliable and zero copy
  - Access to IB via Socket API
- RDS V3
  - supports RDMA read + RDMA write
  - Introduces control messages for asynchronous operation for submit and completion notifications
  - Large data transfers – up to 8 MB
  - Integrated with NUMA IO framework
- RDS diagnostic tools (rds-stress, rds-info, rds-ping)
The InfiniBand Stack

Open Fabrics User Verbs (OFUV)

• API for access to RDMA-capable transport defined by the Open Fabrics Alliance (openfabrics.org)
  – Libraries and utilities for verbs(7) and rdma_cm(7) and kernel subset implementation for RDSv3
  – Provides access mechanism & kernel bypass for Userland

• Provider is assigned IP address via IPoIB
The InfiniBand Stack

Observability

• OpenFabrics Alliance commands & utilities
  – Infiniband diagnostic & observability tools (40+)
  – Eg. perfquery, ibhosts, ibstatus, ibv_devinfo, ibnetdiscover
Future: Exabus - High Speed Network Virtualization

Direct Memory I/O for Java and C++

**Standard Hardware I/O**

- Application
  - Application Buffer
  - TCP IP Transport
  - Kernel

  - 20% Buffer Copies
  - 40% Transport Processing
  - 40% Kernel Context Switches

**Exabus**

- Application
  - Application Buffer

  - Zero Buffer Copy
  - Direct Memory Access
  - Kernel Bypass

**4X Throughput, 6X Lower Latency**
New Exabus Java and C++ APIs

Next-generation of Exalogic performance optimization

• New C++ and Java APIs for maximum application performance
  – Messagebus API accelerates Coherence
  – RDMA API accelerates Tuxedo
  – SDP API accelerates WebLogic

= Coming Soon
Benefits of Engineered Systems

- Full stack testing
- `gettimeofday()` performance 5x
  - Makes tracing much cheaper
- NUMA I/O framework
- Infiniband performance

IB Throughput Improvement, 6 months engineering

+150%
Co-engineered with the Oracle Stack

<table>
<thead>
<tr>
<th></th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU</td>
<td>Fully MT-hot kernel, scales to 100s of cores and 10,000s of HW threads</td>
</tr>
<tr>
<td></td>
<td>Support for Critical Threads features in T4 chip</td>
</tr>
<tr>
<td></td>
<td>JVM support for Solaris scheduling classes, User-level high resolution</td>
</tr>
<tr>
<td></td>
<td>timer support WLS scalability, smt pause() to optimize busy waits in the</td>
</tr>
<tr>
<td></td>
<td>JVM, 5X performance improvement of high-resolution timer</td>
</tr>
<tr>
<td>Memory</td>
<td>Large page support by JVM, T4 2GB pages for Java performance, Preemption</td>
</tr>
<tr>
<td></td>
<td>control NUMA IO framework, Latency-aware kernel memory allocator, NUMA</td>
</tr>
<tr>
<td></td>
<td>optimizations in LDOMs, Intimate Shared Memory (ISM), Dynamic Intimate</td>
</tr>
<tr>
<td></td>
<td>Shared Memory (DISM) Optimized Shared Memory (OSM), NUMA IO framework,</td>
</tr>
<tr>
<td></td>
<td>Latency-aware kernel memory allocator, Large Page support, Fast DB Restart</td>
</tr>
<tr>
<td>File System</td>
<td>Support for Event ports, Userland file system for DB</td>
</tr>
<tr>
<td>I/O</td>
<td>SDP, iPoIB, Receive-side Scaling, LSO for VNICs, Traffic fan-out for EoIB,)</td>
</tr>
<tr>
<td></td>
<td>HA for SDP, Open Fabrics User Verbs, SR-IOV performance scaling, Dynamic</td>
</tr>
<tr>
<td></td>
<td>Reconfiguration for IB HCA,vnet &amp; vswitch performance improvements, uDAPL,</td>
</tr>
<tr>
<td></td>
<td>RDSSv1, RDSSv3, SDP: Support for low-latency InfiniBand protocols, Direct</td>
</tr>
<tr>
<td></td>
<td>I/O with Concurrent writes, Dynamic Reconfiguration for IB HCA</td>
</tr>
<tr>
<td>Security</td>
<td>Integration of JVM with crypto offload engines</td>
</tr>
<tr>
<td></td>
<td>Zones support for EoIB, iPoIB, SDP: Secure isolation, lowest-latency</td>
</tr>
<tr>
<td></td>
<td>virtualization, Integration with Solaris Crypto offload engine, Zones:</td>
</tr>
<tr>
<td></td>
<td>Secure isolation, lowest-latency virtualization</td>
</tr>
</tbody>
</table>
Solaris and Java

Invented, integrated and tested in-house

JVM subsystem instrumentation (DTrace)
- VM Lifecycle Probes
- Thread Lifecycle Probes
- Classloading Probes
- Garbage Collection Probes
- Monitor Probes
- etc.

Security – Java Cryptography Extension+S11
Automatically accelerated security functions,
Solaris PCKS#11 support, evaluating native T4 crypto – bypass JNI

Performance
Sockets Direct Protocol (SDP) enhancements –
RDMA zero-copy data transfers
T4 critical-thread and GC (work in-process)
Transparent large memory page support

Environment aware
The HotSpot JVM and Solaris lgroups
(latency groups) helps ensure that Java
objects are placed in local memory on
NUMA machines
Smarter Threads
T4 and Solaris Dispatcher coordinate to enable dynamic threads. Applications run 5x faster

Effortless Crypto
Automatically accelerated Java, application and Oracle Database encryption. 4x faster OpenSSL vs IBM

Big Pipes
Deliver dedicated bandwidth to network intensive services with 10GigE network pipes and Oracle Solaris 11 network virtualization

Power Smart
Automatic conservation of threads, cores and memory by the Solaris Power Dispatcher to cut your power expenses
## Optimized for x86

Co-engineered with Intel

| Performance          | Optimized `memcpy`, `memset`, string routines, `gettimeofday()` 5x faster  
|                      | SSE3/4.1/4.2, AVX  
|                      | Performance counters for all new CPUs, `latencytop`  
|                      | gcc 4.5.2  
| Power Management     | `powertop`, `poweradm`  
|                      | Full C-State and P-State support  
|                      | Turbo Boost observability, `RAPL Power Meter`  
| Security             | Hardware crypto acceleration and optimization: AES-NI, SHA1 w/SSE3, RC4…  
|                      | Trusted Platform Module (TPM) driver and keystore support  
|                      | Executable page removal via NX bit  
| I/O                  | Storage: SCU support, 30%+ improvement for large I/Os using IOMMU  
|                      | Graphics: GRFX, DRM, AGP, VESA console, virtual consoles, graphical startup  
|                      | Network: SR-IOV, `ixgbe`, `e1000g`, wireless, Zero-copy TCP with I/OAT  
|                      | PCIe hotplug, CPU microcode update  
| Fault Tolerance      | Extensive hardware-specific fault detection and recovery with FMA, MCA support, `CPU`, `memory`, and I/O retire  
|                      | Auto-discovery of memory fault topologies |
For More Information / Try Out Today

• Product overview and download
  – oracle.com/solaris

• Oracle Technology Network
  – oracle.com/technetwork/server-storage/solaris11

• System administrators community
  – oracle.com/technetwork/systems

@ORCL_Solaris
facebook.com/oraclesolaris
Oracle Solaris Insider
Hardware and Software

ORACLE

Engineered to Work Together
The Unique Oracle Advantage
Everything Just Works Together