UCS Invicta:
A New Generation of Storage Performance

Mazen Abou Najm
DC Consulting Systems Engineer
HDDs Aren’t Designed For High Performance
Disk 101

<table>
<thead>
<tr>
<th>Speed</th>
<th>Low Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latency in Seconds</td>
<td>0.001 (milliseconds)</td>
</tr>
<tr>
<td>Transfer rate(s) MB/s</td>
<td>10s</td>
</tr>
<tr>
<td>Write /Read operations per Second (IOPS)</td>
<td>100s</td>
</tr>
</tbody>
</table>

- Can’t spin faster (200 IOPS/Drive)
- Can’t seek faster (6-8 ms latency)
- Only Performance Option is to overprovision or short stroking
- Power, Cooling, & Rack-Space waste
- Escalating costs

- Design
  - Mechanical
  - Motors & Spindles
  - High Energy consumption
The Trade-Offs

Data Persistence

Protection

Performance

$\text{COST}$
Flash Is Designed To Deliver Higher Performance & Lower Operating Costs

<table>
<thead>
<tr>
<th></th>
<th>Low Performance</th>
<th>High Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Est</strong></td>
<td>1956</td>
<td>1980</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Hard Disk Drive</th>
<th>Flash Drive</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Latency in Seconds</strong></td>
<td>0.001 (milliseconds)</td>
<td>0.000001 (microseconds)</td>
</tr>
<tr>
<td><strong>Transfer rate(s) MB/s</strong></td>
<td>10s</td>
<td>100s</td>
</tr>
<tr>
<td><strong>Write /Read operations per Second ( IOPS)</strong></td>
<td>100s</td>
<td>1000s</td>
</tr>
</tbody>
</table>

**Design**
- **Mechanical**
- ** Motors & Spindles**
- **High Energy consumption**
- **Silicon**
- **Integrated Circuit**
- **Low Energy Consumption**
Disk 101 - RAID & IOPS

<table>
<thead>
<tr>
<th>RAID</th>
<th>READ Penalty</th>
<th>Write Penalty</th>
<th>Capacity Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1 &amp; 10</td>
<td>1</td>
<td>2</td>
<td>#Disks/2</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>4</td>
<td>#Disks-1 Disk</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>6</td>
<td>#Disks-2 Disks</td>
</tr>
</tbody>
</table>

BE IOPS Required = ((FE IOPS x %READ)+(FE IOPS x %Write) x RAID Write Penalty)

Example

Application requires 100000 IOPS with 50% Read and 50% Write and you’re using RAID5 & 15K Drives with 200 IOPS

((100000x50%)+(100000x50%)*4)=250000 BE IOPS Required (1250 Disks Required)
Disk Provisioning For Performance

Reference Architecture for 1,000 Desktops

- 3 types of drives, 3 types of RAID
- 41 15K HDDs
- 25 7.2K HDDs
- 3 Flash Drives

1,000 Persistent Desktops will require:
- < 10TB of capacity
- ~80K backend IOPS

TOTAL IOPS: 114,950
TOTAL CAPACITY: 63.2 TB
Flash 101

<table>
<thead>
<tr>
<th>Speed</th>
<th>High Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latency in Seconds</td>
<td>0.000001 (microseconds)</td>
</tr>
<tr>
<td>Transfer rate(s) MB/s</td>
<td>100s</td>
</tr>
<tr>
<td>Write /Read operations per Second ( IOPS)</td>
<td>1000s</td>
</tr>
</tbody>
</table>

Design

- Silicon
- Integrated Circuit
- Low Energy Consumption

- Inherent strengths
  - Low latency
  - High read speed
  - Durability
  - Low power
  - Small footprint

- Inherent weaknesses
  - High cost
  - Very poor write speed
  - Low endurance
Flash 101 - Glossary

- NAND Page – Normally 4K in size and multiple pages form Erase Blocks
- Erase Block – Collection of NAND Pages normally 1 or 2MB in size
- Page Erase – The process required to allow an incoming write to a NAND Page in Erase Block
- Write Amplification - Actual amount of physical information written is a multiple of the logical amount intended to be written.
- Wear Leveling - Arranging data so that erasures and re-writes are distributed evenly across the medium.
- GC – Garbage Collection is a process of reclaiming NAND Pages to create free or empty Erase Blocks
- SSD – Solid State Drives (NAND on Drive)
## Flash 101 – Different Types

<table>
<thead>
<tr>
<th>States</th>
<th>Erase Cycles</th>
<th>Max $t_{Prog}$</th>
<th>$t_{R}$</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLC</td>
<td>2</td>
<td>100,000</td>
<td>700us</td>
<td>$$$</td>
</tr>
<tr>
<td>MLC</td>
<td>4</td>
<td>3,000–10,000</td>
<td>1200us</td>
<td>$$</td>
</tr>
<tr>
<td>TLC</td>
<td>8</td>
<td>1,000</td>
<td>2000us</td>
<td>$</td>
</tr>
</tbody>
</table>

SLC = Single Level Cell  
MLC = Multi Level Cell  
TLC = Triple Level Cell  

$^t_{Prog}$ – Time to transfer contents of data register to flash  
$^t_{R}$ – Time to transfer contents of 1 flash page to data register
Flash 101 - Flash Write Process

1. Erase Block contents are read to a buffer.
2. Erase Block is erased (aka, “flashed”).
3. Buffer is written back with previous data and any changed or new blocks – including zeroes.
If block write issued begins at the start of a physical sector on storage, no additional writes and no partial writes.

If each block write issued does not begin at the start of a physical sector on storage, misalignment results in additional writes as the last block overlaps into a new sector – and consequently results in at least one partial write.

Results in Latency and Endurance Problems
Flash Provisioning for Performance

- 41 15K HDDs
- 25 7.2K HDDs
- 3 Flash Drives

Total IOPS: 114,950
Total Capacity: 63.2 TB

- 200,000 IOPS
- 64 TB
Introducing the Cisco UCS Invicta Series

UCS Invicta Scaling System

Using Invicta OS 5.0.0

First release:
Up to 1.3 Million IOPS
Up to 13.2 GB/s Bandwidth
Up to 240TB RAW

UCS Invicta Appliance

First release:
250,000 IOPS
1.9 GB/s Bandwidth
Up to 24 TB RAW

- Scalability
- Modularity
- Application Acceleration
- Data Optimization
- Multiple Workloads
- Tuning-Free Performance

Cisco Connect, Riyadh, Saudi Arabia, April 29-30, 2014
## Introducing the Cisco UCS Invicta Series

<table>
<thead>
<tr>
<th>Workload Acceleration</th>
<th>Data Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Appliance</strong></td>
<td><strong>Silicon Node</strong></td>
</tr>
<tr>
<td>Bandwidth (GB/s)</td>
<td>1.9</td>
</tr>
<tr>
<td>IOPS</td>
<td>250,000</td>
</tr>
<tr>
<td>Latency (Microseconds)</td>
<td>&lt;100</td>
</tr>
<tr>
<td>Size</td>
<td>2 RU</td>
</tr>
<tr>
<td>Max Capacity (TB)</td>
<td>24</td>
</tr>
</tbody>
</table>

* Effective Capacity

_Cisco Connect, Riyadh, Saudi Arabia, April 29-30, 2014_
# Introducing the Cisco UCS Invicta Series

<table>
<thead>
<tr>
<th>Silicon Router</th>
<th>Silicon Node</th>
</tr>
</thead>
<tbody>
<tr>
<td>Host Connectivity</td>
<td>Flash Memory Devices</td>
</tr>
<tr>
<td>CPU, Memory &amp; RACERUNNER OS</td>
<td>CPU, Memory &amp; RACERUNNER OS</td>
</tr>
</tbody>
</table>

### Switched Fabric

<table>
<thead>
<tr>
<th>Storage Functions</th>
<th>Storage Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Host Presentation</td>
<td>• Flash Management</td>
</tr>
<tr>
<td>• Mirroring</td>
<td>• Volume Management</td>
</tr>
<tr>
<td>• Replication</td>
<td>• RAID</td>
</tr>
<tr>
<td>• Snapshots</td>
<td>• De-Duplication</td>
</tr>
<tr>
<td>• Reporting</td>
<td>• Thin Provisioning</td>
</tr>
<tr>
<td>• Node Grouping</td>
<td>• Power-Fail Data Protection</td>
</tr>
<tr>
<td>• Striping</td>
<td></td>
</tr>
</tbody>
</table>
Introducing the Cisco UCS Invicta Series

- **Scale Up**
- **Scale Out**
- **Scale Up/Out**

<table>
<thead>
<tr>
<th>Add Capacity</th>
<th>Add Nodes</th>
<th>Add Routers &amp; Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Add capacity to the Storage System</td>
<td>Add Storage Systems to a networked pool</td>
<td>Add Storage Systems to Networked Routed Infrastructure</td>
</tr>
<tr>
<td>Performance is fixed</td>
<td>Data is distributed across nodes to increase aggregate performance</td>
<td>Routers organize pools consisting of one or more storage systems</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Distributing data across storage systems increases aggregate IOPS, throughput &amp; capacity</td>
</tr>
</tbody>
</table>
Introducing the Cisco UCS Invicta Series

<table>
<thead>
<tr>
<th>Add Capacity</th>
<th>Add Nodes</th>
<th>Add Routers &amp; Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scale Up Array</strong></td>
<td><strong>Scale Out Nodes</strong></td>
<td><strong>Scale Up/Out</strong></td>
</tr>
<tr>
<td>Storage Management (control plane)</td>
<td>Centralized</td>
<td>Duplicated/Distributed</td>
</tr>
<tr>
<td>SAN Port Consumption</td>
<td>Controllers (Few)</td>
<td>Nodes (Many)</td>
</tr>
<tr>
<td>Operating System</td>
<td>Centralized</td>
<td>Distributed</td>
</tr>
<tr>
<td>Performance Scaling</td>
<td>Fixed</td>
<td>N+1</td>
</tr>
<tr>
<td>Device Connectivity</td>
<td>Point-to-Point</td>
<td>Peer Network</td>
</tr>
<tr>
<td>Device Capabilities</td>
<td>Storage Shelves</td>
<td>Uniform Nodes</td>
</tr>
<tr>
<td>Data Placement</td>
<td>Constrained</td>
<td>Distributed</td>
</tr>
<tr>
<td>Data Protection</td>
<td>RAID</td>
<td>Non-RAID</td>
</tr>
<tr>
<td>Data Replication</td>
<td>Controller Function</td>
<td>Generally No</td>
</tr>
</tbody>
</table>
Introducing the Cisco UCS Invicta Series

Using Invicta OS 5.0

Cisco Connect, Riyadh, Saudi Arabia, April 29-30, 2014
# Introducing the Cisco UCS Invicta Series

<table>
<thead>
<tr>
<th>Feature</th>
<th>Invicta OS</th>
<th>Scale Up / Scale out Architecture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appliance</td>
<td>5.0 and up</td>
<td>5.0</td>
</tr>
<tr>
<td>Symmetric Read/Writes</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>RAID Protection</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Asynchronous Replication</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Snapshots – Copy on Write</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Mirroring</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Web Based UI/API</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Role Based Access Controls</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Data Reduction Option</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Thin Provisioning</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>iSCSI, Fibre Channel</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Enhanced Data Protection</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>VAAI / vCenter Support</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Ethernet, Fibre Channel</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Increase Performance and Capacity</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Switched Fabric</td>
<td>7 Nodes &amp; Up</td>
<td>7 Nodes &amp; Up</td>
</tr>
<tr>
<td>Max. Routers</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Max Nodes</td>
<td>10</td>
<td>30</td>
</tr>
</tbody>
</table>

Cisco Connect, Riyadh, Saudi Arabia, April 29-30, 2014
Introducing the Cisco UCS Invicta Series

Using Invicta OS 5.1

Cisco Connect, Riyadh, Saudi Arabia, April 29-30, 2014
Introducing the Cisco UCS Invicta Series

2 Routers & 30 Nodes
• Throughput 14 GB/s
• IOPs 4,050,000
• Raw Capacity 720TB

3 Routers & 27 nodes
• Throughput 21 GB/s
• IOPs 3,645,000
• Capacity 684TB

4 Routers & 24 Nodes
• Throughput 28 GB/s
• IOPs 3,240,000
• Capacity 576TB

5 Routers & 21 Nodes
• Throughput 35 GB/s
• IOPs 2,835,000
• Capacity 504TB

6 Routers & 18 Nodes
• Throughput 42 GB/s
• IOPs 2,430,000
• Capacity 432TB
UCS Invicta – Under The Covers
Cisco UCS Invicta Series – Under The Covers
Receive
• Data blocks of various sizes arrive from Hosts.

Protect
• Data Blocks are stored in the power loss buffer and passed onto the Block Translation Layer (BTL)

Optimize
• The Block Translation Layer aggregates and sizes Data Blocks for the RAID Layer and Flash Media

Commit
• BTL Optimized Data is flushed across the RAID stripe and written to Flash media concurrently.
Cisco UCS Invicta Series – Under The Covers

Inbound data blocks

Fill

Write

Cache

Block Translation Layer

Optimize

Write

Cisco Connect, Riyadh, Saudi Arabia, April 29-30, 2014
Cisco UCS Invicta Series – Under The Covers

Treats NAND Flash like NAND not like disk
- NEVER writes less than an entire Erase block
- Smaller writes are padded to the EB boundary
- Writes are acknowledged to initiator immediately after being recorded into NV memory

Short writes (smaller than chunk) require reading all elements to calculate parity
- Worst case in a 24 drive system is 21 reads to do a SINGLE write operation

UCS Invicta BTL *NEVER* does this
- WBs are multiples of stripe size and EB size
- Allows for less wear AND better $/GB
Cisco UCS Invicta Series – Under The Covers

Cisco UCS Invicta has integrated data de-duplication
  • Configured at order time
  • Up to 10:1 overcommit ratio

High performance and IN-LINE
  • ~ 200K IOPS READ
  • ~ 160K IOPS Write
  • Integrated into VAAI (XCOPY and WRITE-SAME)

Deduplication Uses UCS Invicta BTL Hashes
  • Hash function is high performance (MM3)
  • Hash is compared to all existing hashes in Memory
  • Media read verifies the duplication
  • A miss-match forces a unique store
  • A full match stores a pointer
Cisco UCS Invicta Series – Under The Covers

Write Tests:

<table>
<thead>
<tr>
<th>Block Size</th>
<th>1 thread</th>
<th>10 threads</th>
<th>40 threads</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IOPS</td>
<td>BW</td>
<td>IOPS</td>
</tr>
<tr>
<td>512B</td>
<td>80583</td>
<td>39.3M</td>
<td>82024</td>
</tr>
<tr>
<td>1K</td>
<td>77820</td>
<td>75.9M</td>
<td>83648</td>
</tr>
<tr>
<td>2K</td>
<td>80463</td>
<td>157.1M</td>
<td>83710</td>
</tr>
<tr>
<td>4K</td>
<td>142634</td>
<td>557.1M</td>
<td>263742</td>
</tr>
<tr>
<td>8K</td>
<td>98430</td>
<td>768.9M</td>
<td>164292</td>
</tr>
<tr>
<td>16K</td>
<td>64868</td>
<td>1013.5M</td>
<td>85322</td>
</tr>
<tr>
<td>32K</td>
<td>32530</td>
<td>1016.5M</td>
<td>43143</td>
</tr>
<tr>
<td>64K</td>
<td>14262</td>
<td>891.3M</td>
<td>20344</td>
</tr>
<tr>
<td>128K</td>
<td>8056</td>
<td>1007.0M</td>
<td>9292</td>
</tr>
<tr>
<td>256K</td>
<td>4716</td>
<td>1179.1M</td>
<td>4542</td>
</tr>
<tr>
<td>512K</td>
<td>2545</td>
<td>1272.5M</td>
<td>2365</td>
</tr>
<tr>
<td>1M</td>
<td>1041</td>
<td>1041.0M</td>
<td>1378</td>
</tr>
<tr>
<td>2M</td>
<td>626</td>
<td>1252.7M</td>
<td>608</td>
</tr>
<tr>
<td>4M</td>
<td>315</td>
<td>1260.0M</td>
<td>331</td>
</tr>
</tbody>
</table>

4K 100% Random Writes

Thursday, Jan 16, 2014 11:06:26
MB/s: 1094.356
IOPS: 273589
Latency: 20 μs
Silos vs. Solution

Cisco Connect, Riyadh, Saudi Arabia, April 29-30, 2014
UCS Introduces Flash Memory

NEXT GEN UNIFIED COMPUTING
Integration of Solid-State Memory Systems into the UCS Fabric

UCS Founding Principles

| Application Centricity | Operational Simplicity | Platform for IT Innovation |

UCS Invicta Series Solid-State Systems

Address new data velocity and scale requirements
Integrate application acceleration into the computing domain

Cisco Connect, Riyadh, Saudi Arabia, April 29-30, 2014
Flash Memory Brings Application Acceleration into Modular Infrastructure with Unified Resource Management

UCS becomes the Fast Lane for Application Workloads
Recap - UCS Invicta Series

UCS Invicta Series
- Part of the UCS Architecture
- Modular Architecture
- UCS Management Integration
- Future Fabric Integration

Appliance
- Invicta OS
- Media Management, Data Protection & Data Reduction
- Balanced Write/Read Performance
- Converts into Scaling System Node

Scaling System
- Invicta OS
- Router Services
- Media Management, Data Protection & Data Reduction
- Efficiency of Scale Up with the Power of Scale Out
- Supports Multiple Workloads
UCS Invicta Series
Bringing Flash Memory into the UCS Architecture

Invicta OS
V5.x
• Flash media management
• Data protection
• De-Duplication
• UCS Director Support
• FCoE

UCS Invicta Series
Appliance
Scaling System
Storage Blade
-----
• Based upon
  • UCS C
  • UCS B

Network
VIC
Fabric Interconnect
-----
Unified Fabric
• Lower latency
• Higher bandwidth

Management
UCS Director
UCS Manager
-----
• Orchestration
• Policies
• Service Profiles
• Self-service
• Tasks
• Workflows

SAN/LAN
B Series
C Series
UCS
Invicta Series

SAN/LAN Fabric Interconnect

Cisco Connect, Riyadh, Saudi Arabia, April 29-30, 2014

Items in green are completed or in progress