cisco

Cisco Connect

Riyadh, Saudi Arabia

April 29-30, 2014

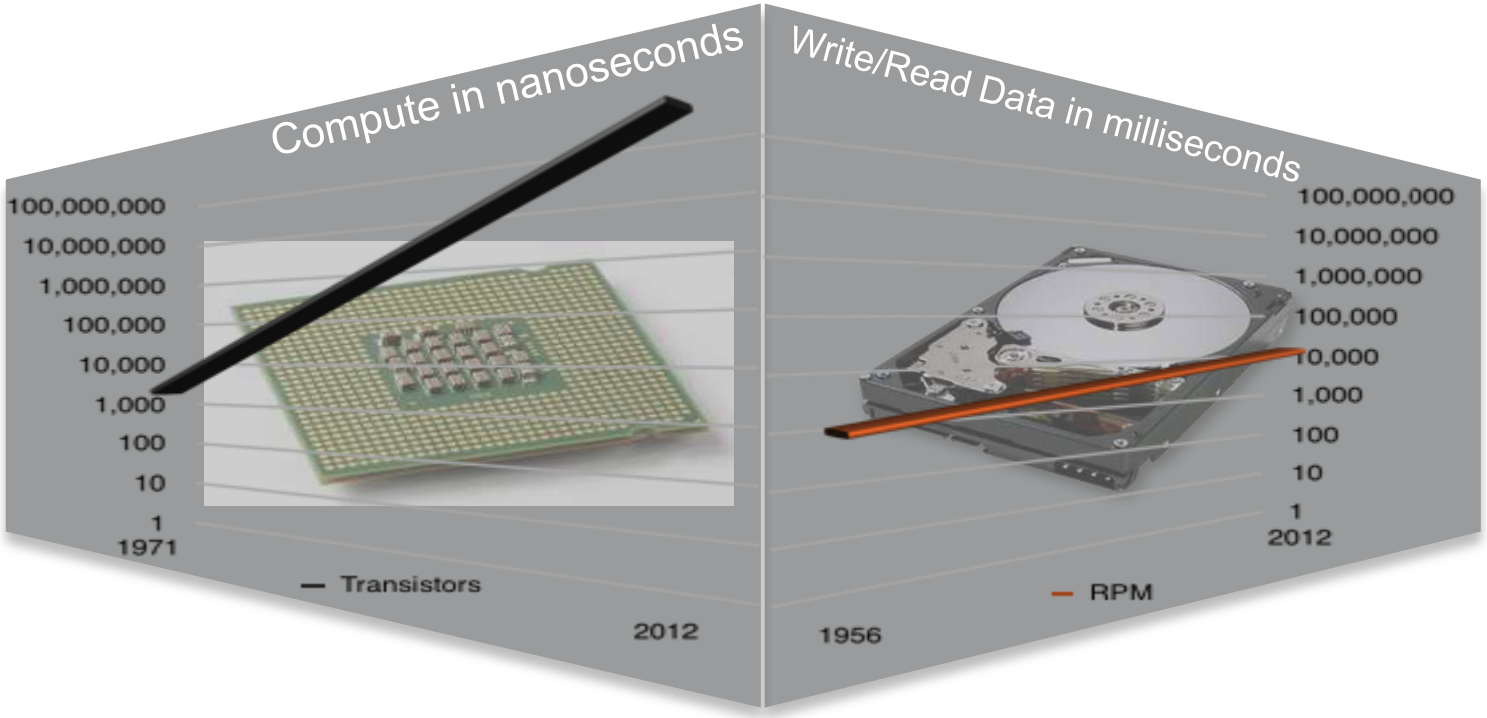
TOMORROW starts here.

UCS Invicta:
A New Generation of Storage Performance

Mazen Abou Najm

DC Consulting Systems Engineer

HDDs Aren’t Designed For High Performance

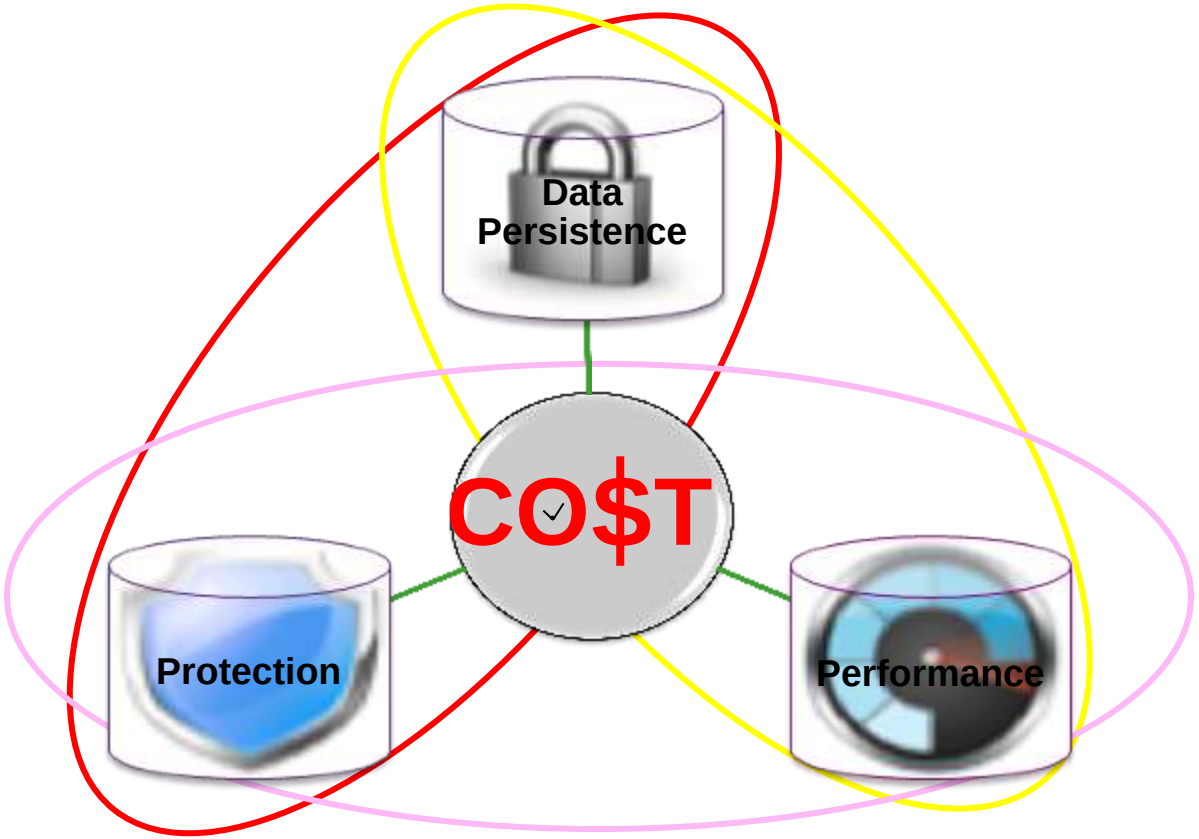


Disk 101

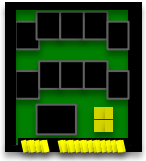
		Low Performance
Speed	Latency in Seconds	0.001 (milliseconds)
	Transfer rate(s) MB/s	10s
	Write /Read operations per Second (IOPS)	100s
Design		Mechanical
		Motors & Spindles
		High Energy consumption

- 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

The Trade-Offs



Flash Is Designed To Deliver Higher Performance & Lower Operating Costs

		Low Performance	High Performance
		Est 1956	Est 1980
			
		Hard Disk Drive	Flash Drive
Speed	Latency in Seconds	0.001 (milliseconds)	0.000001 (microseconds)
	Transfer rate(s) MB/s	10s	100s
	Write /Read operations per Second (IOPS)	100s	1000s
Design		Mechanical	Silicon
		Motors & Spindles	Integrated Circuit
		High Energy consumption	Low Energy Consumption

Disk 101 - RAID & IOPS

RAID	READ Penalty	Write Penalty	Capacity Impact
0	1	1	0
1 & 10	1	2	#Disks/2
5	1	4	#Disks-1 Disk
6	1	6	#Disks-2 Disks

Disk Speed	IOPS
7,200 RPM	75-100
10,000 RPM	125-150
15,000 RPM	175-210

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%)x4)=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

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

TOTAL IOPS: 114,950
TOTAL CAPACITY: 63.2 TB

Flash 101

		High Performance
Speed	Latency in Seconds	0.000001 (microseconds)
	Transfer rate(s) MB/s	100s
	Write /Read operations per Second (IOPS)	1000s
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

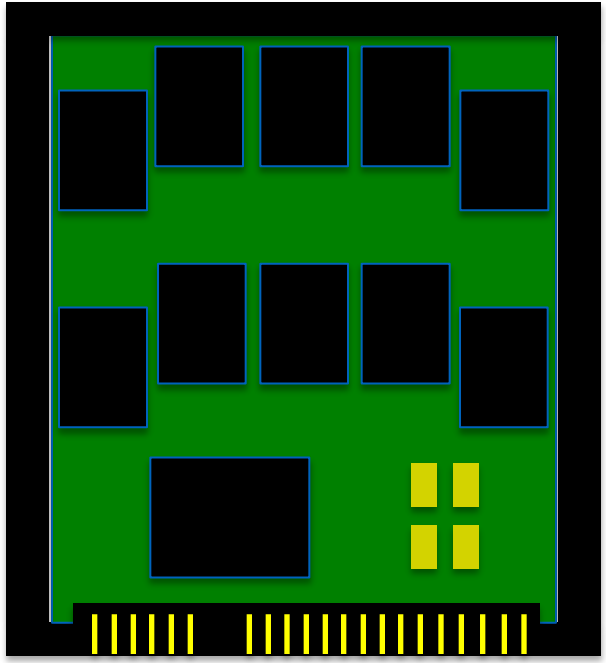
	States	Erase Cycles	Max ^{t}Prog	^{t}R	Cost
SLC	2	100,000	700us	25us	$$$
MLC	4	3,000-10,000	1200us	50us	$$
TLC	8	1,000	2000us	100us	$

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

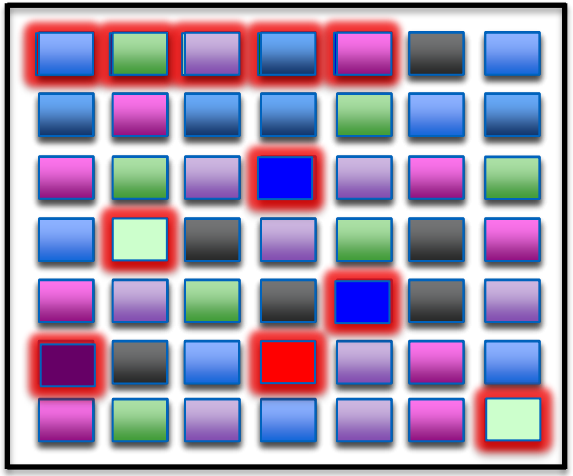
^{t}Prog – Time to transfer contents of data register to flash

^{t}R – Time to transfer contents of 1 flash page to data register

Flash 101 - Flash Write Process

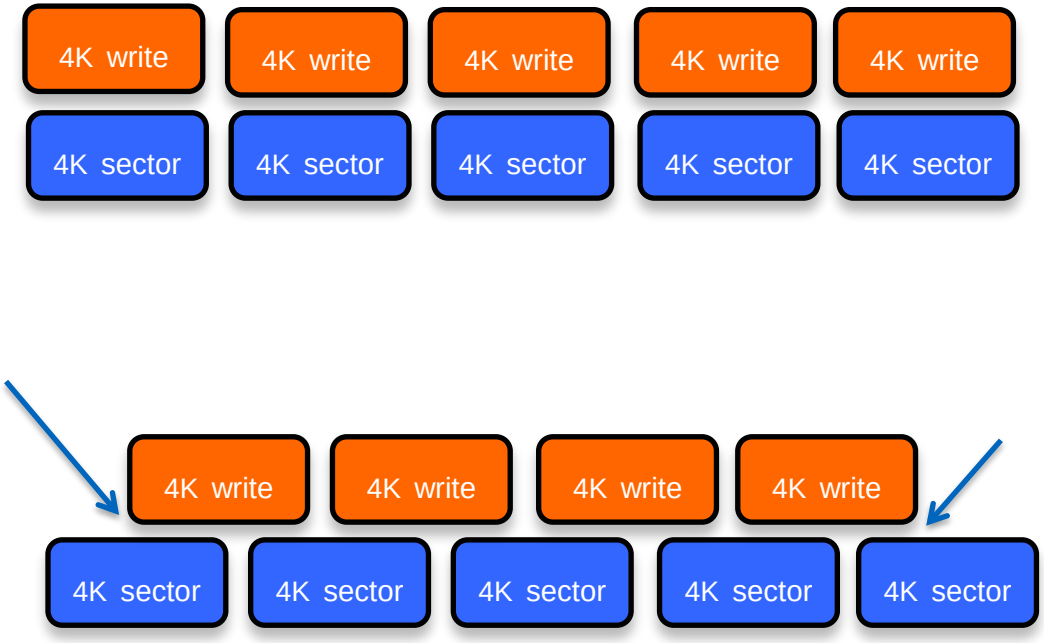


2MB Erase Block



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.

Flash 101 - Write Alignment

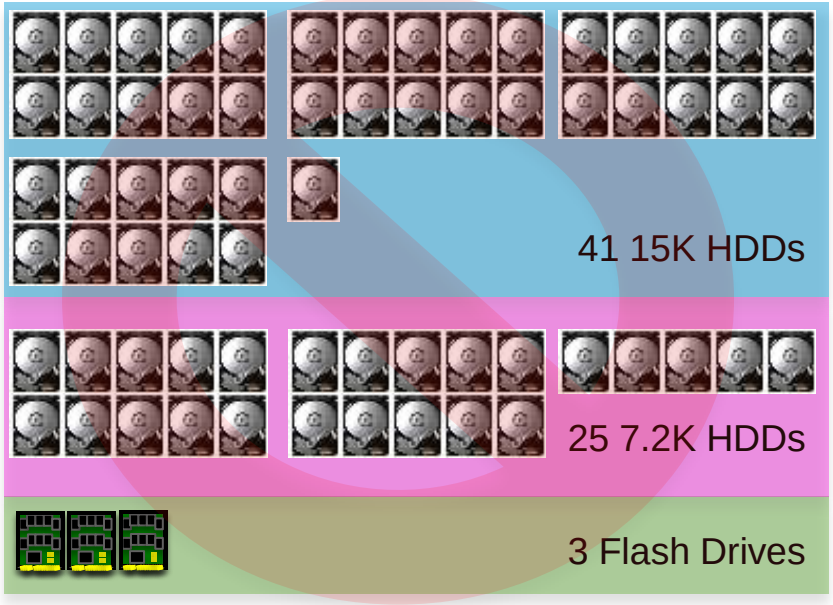


If block write issued begins at the start of a physical sector on storage, no additional writes and no partial writes

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



114,950 IOPS 63.2 TB

200,000 IOPS 64 TB

Cisco UCS Invicta

Introducing the Cisco UCS Invicta Series

UCS Invicta Scaling System

First release:

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

Using Invicta OS 5.0.0

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

Introducing the Cisco UCS Invicta Series

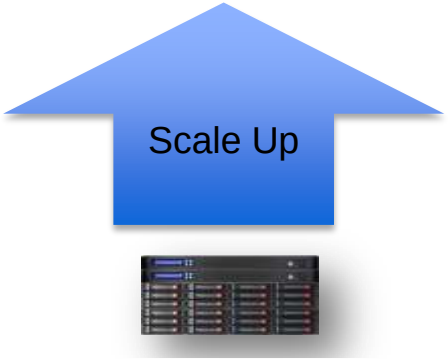
	Workload Acceleration		Data Reduction	
				
	Appliance	Silicon Node	Appliance	Silicon Node
Bandwidth (GB/s)	1.9	1.5	1.5	1.2
IOPS	250,000	200,000	200,000	165,000
Latency (Microseconds)	<100	<200	<100	<200
Size	2 RU		2 RU	
Max Capacity (TB)	24		64*	

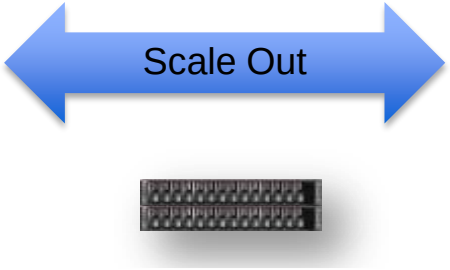
* Effective Capacity

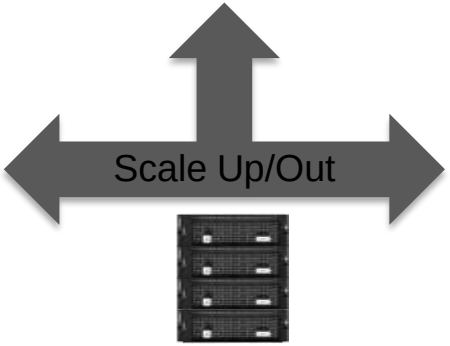
Introducing the Cisco UCS Invicta Series

Silicon Router	Silicon Node
Host Connectivity	Flash Memory Devices
CPU, Memory & RACERUNNER OS	CPU, Memory & RACERUNNER OS
Switched Fabric	
Storage Functions • Host Presentation • Mirroring • Replication • Snapshots • Reporting • Node Grouping • Striping	Storage Functions • Flash Management • Volume Management • RAID • De-Duplication • Thin Provisioning • Power-Fail Data Protection

Introducing the Cisco UCS Invicta Series

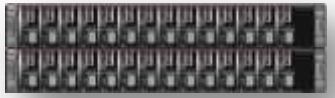






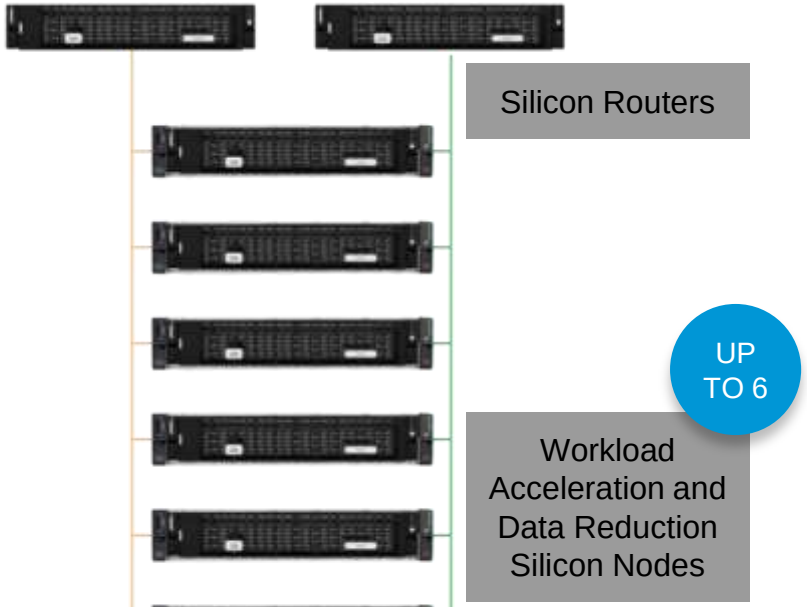
Add Capacity	Add Nodes	Add Routers & Nodes
Add capacity to the Storage System	Add Storage Systems to a networked pool	Add Storage Systems to Networked Routed Infrastructure
Performance is fixed	Data is distributed across nodes to increase aggregate performance	Routers organize pools consisting of one or more storage systems
		Distributing data across storage systems increases aggregate IOPS, throughput & capacity

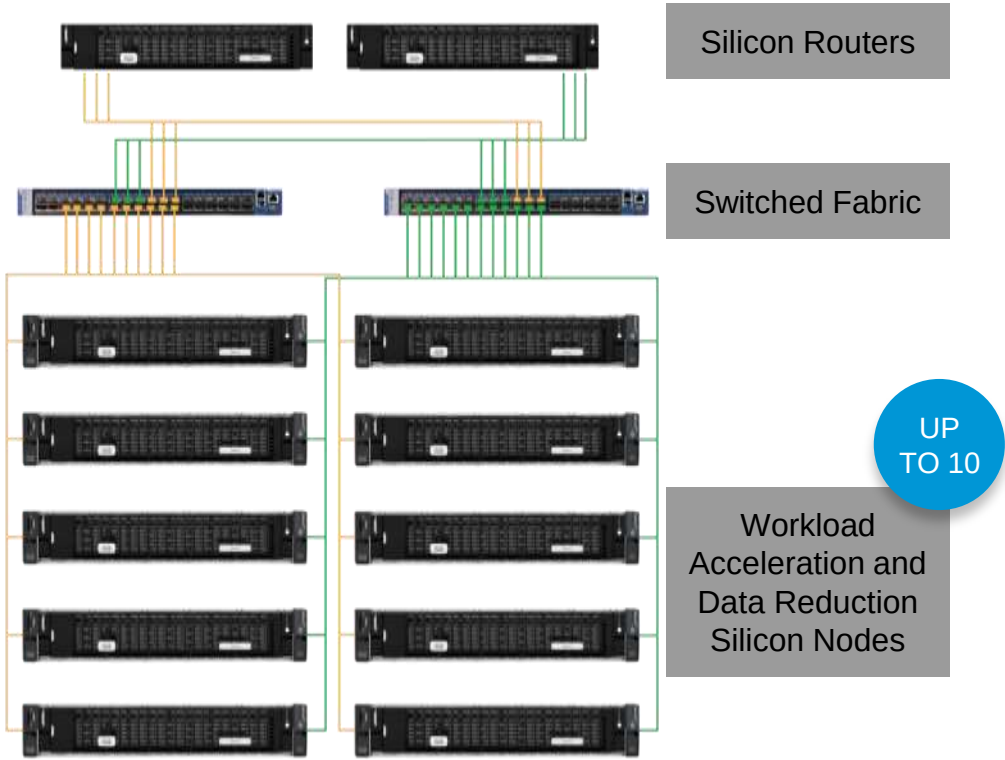
Introducing the Cisco UCS Invicta Series



	Add Capacity	Add Nodes	Add Routers & Nodes
	Scale Up Array	Scale Out Nodes	Scale Up/Out
Storage Management (control plane)	Centralized	Duplicated/Distributed	Separated/Distributed
SAN Port Consumption	Controllers (Few)	Nodes (Many)	Routers (Few)
Operating System	Centralized	Distributed	Distributed
Performance Scaling	Fixed	N+1	RxN+1
Device Connectivity	Point-to-Point	Peer Network	Routed Network
Device Capabilities	Storage Shelves	Uniform Nodes	Flexible Nodes
Data Placement	Constrained	Distributed	Constrained or Distributed
Data Protection	RAID	Non-RAID	RAID
Data Replication	Controller Function	Generally No	Router Function

Introducing the Cisco UCS Invicta Series



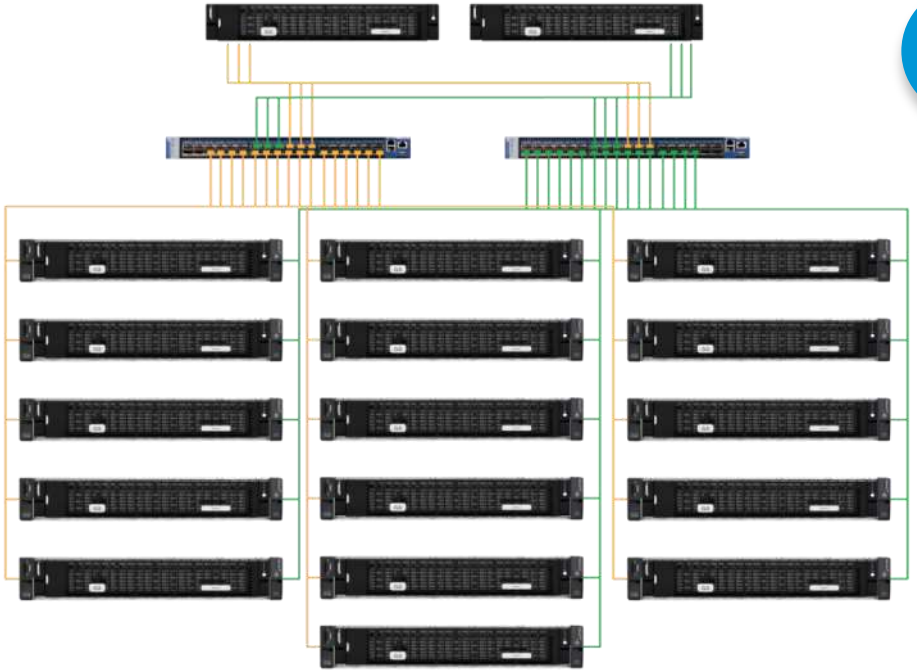


Using Invicta OS 5.0

Introducing the Cisco UCS Invicta Series

	Appliance	Scale Up / Scale out Architecture	
Invicta OS	5.0 and up	5.0	5.1.0
Symmetric Read/Writes	✔	✔	✔
RAID Protection	✔	✔	✔
Asynchronous Replication	✔	✔	✔
Snapshots – Copy on Write	✔	✔	✔
Mirroring		✔	✔
Web Based UI/API	✔	✔	✔
Role Based Access Controls	✔	✔	✔
Data Reduction Option	✔	✔	✔
Thin Provisioning	✔	✔	✔
iSCSI, Fibre Channel	✔	✔	✔
Enhanced Data Protection	✔	✔	✔
VAAI / vCenter Support	✔	✔	✔
Ethernet, Fibre Channel	✔	✔	✔
Increase Performance and Capacity		✔	✔
Switched Fabric		7 Nodes & Up	7 Nodes & Up
Max. Routers		2	6
Max Nodes		10	30

Introducing the Cisco UCS Invicta Series



UP TO 6

Silicon Routers

Switched Fabric

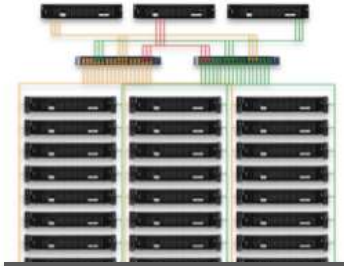
Workload Acceleration and Data Reduction Silicon Nodes

UP TO 30

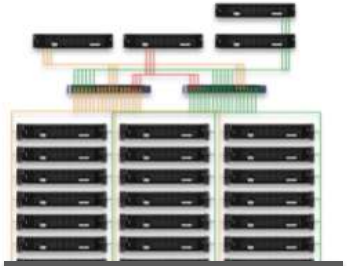
Using Invicta OS 5.1

Introducing the Cisco UCS Invicta Series

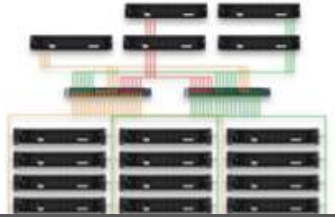
2 Routers & 30 Nodes
- Throughput 14 GB/s
- IOPs 4,050,000
- Raw Capacity 720TB



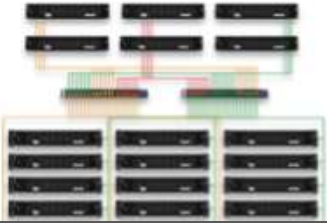
3 Routers & 27 nodes
- Throughput 21 GB/s
- IOPS 3,645,000
- Capacity 684TB



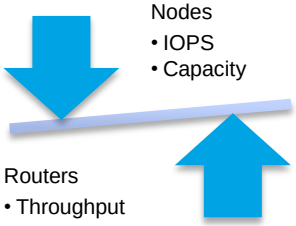
4 Routers & 24 Nodes
- Throughput 28 GB/s
- Capacity 576TB
- IOPS 3,240,000



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

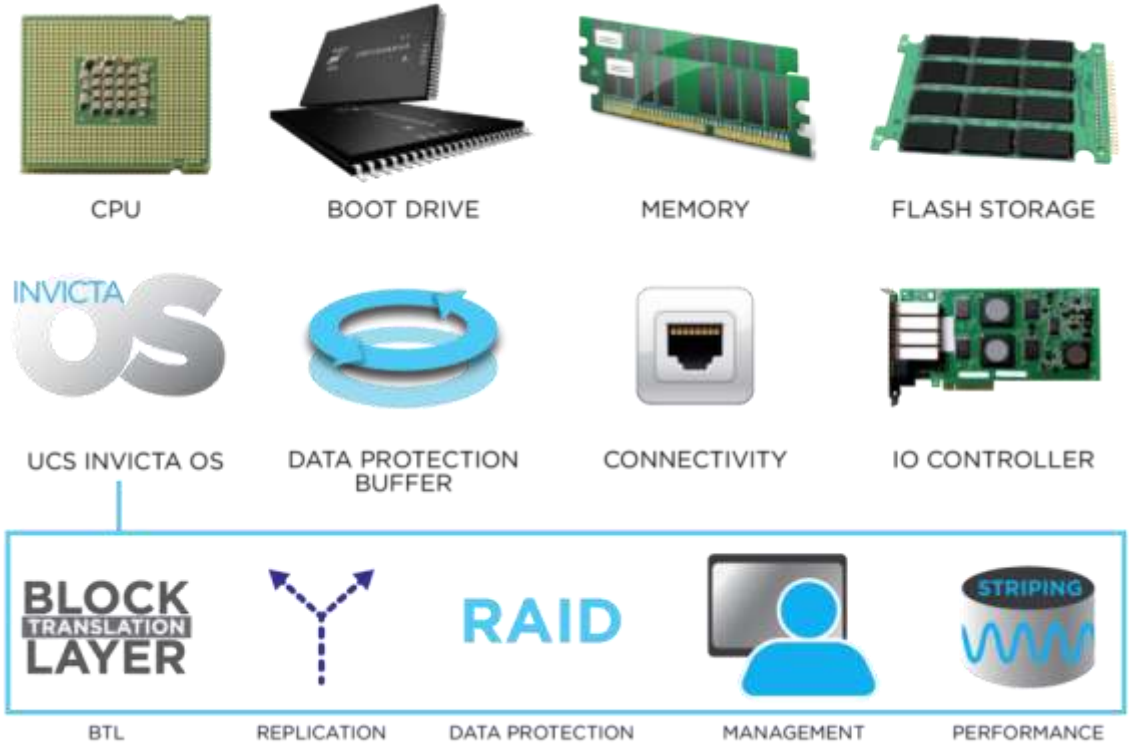


6 Routers & 18 Nodes
- Throughput 42 GB/s
- IOPS 2,430,000
- Capacity 432TB



UCS Invicta – Under The Covers

Cisco UCS Invicta Series – Under The Covers



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)

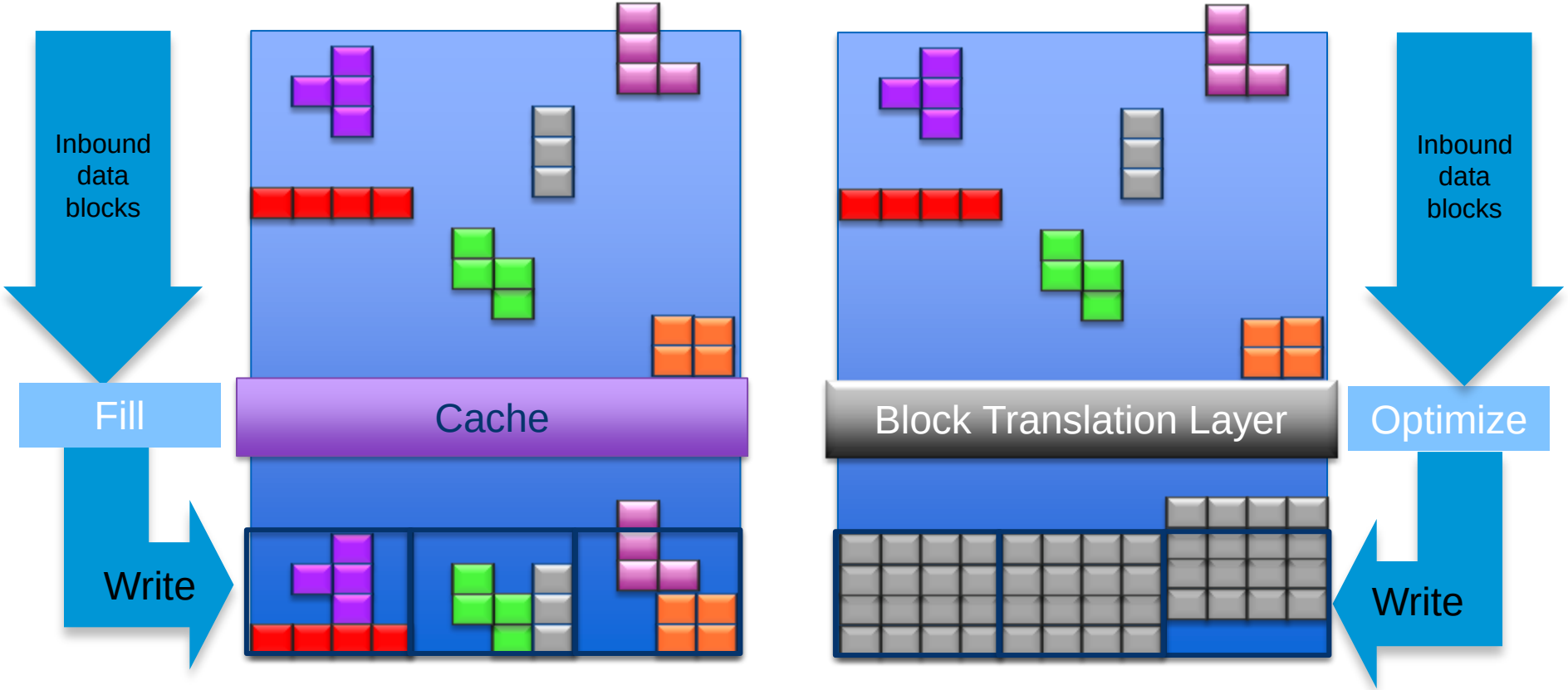
Optimize

- 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



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

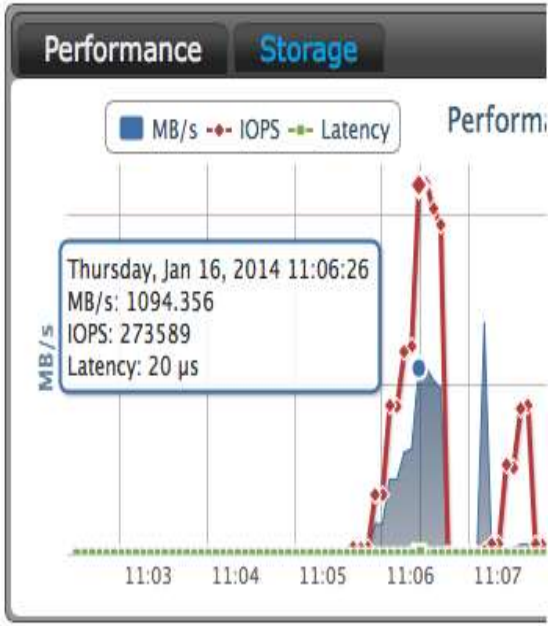
High performance and IN-LINE

- ~ 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

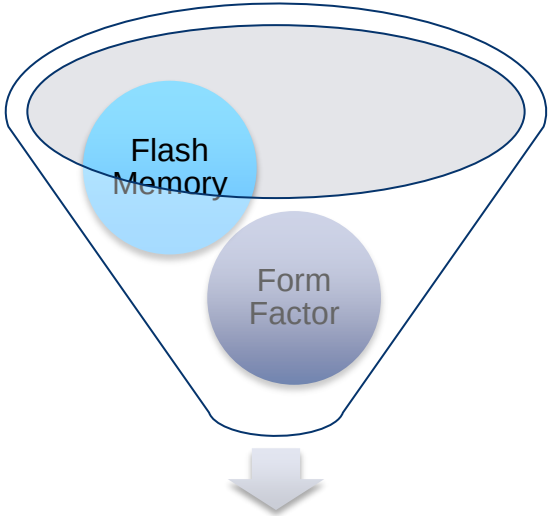


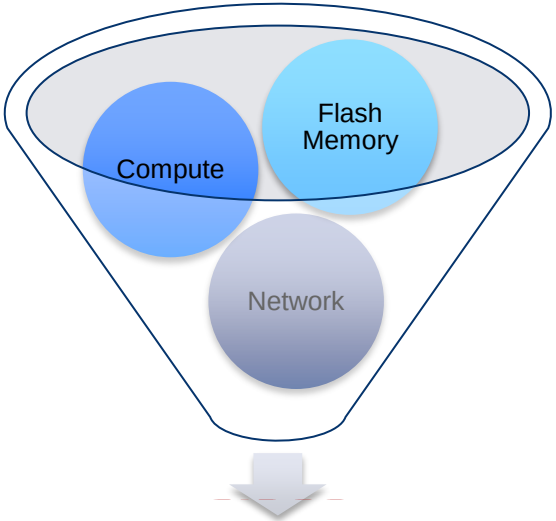
Write Tests:

Block Size	1 thread IOPS	1 thread BW	10 threads IOPS	10 threads BW	40 threads IOPS	40 threads BW
512B	80583	39.3M	82024	40.0M	81011	39.5M
1K	77820	75.9M	83648	81.6M	79474	77.6M
2K	80463	157.1M	83710	163.4M	80874	157.9M
4K	142634	557.1M	263742	1030.2M	192401	751.5M
8K	98430	768.9M	164292	1283.5M	145863	1139.5M
16K	64868	1013.5M	85322	1333.1M	74077	1157.4M
32K	32530	1016.5M	43143	1348.2M	36605	1143.9M
64K	14262	891.3M	20344	1271.5M	15281	955.0M
128K	8056	1007.0M	9292	1161.5M	7212	901.5M
256K	4716	1179.1M	4542	1135.5M	3413	853.2M
512K	2545	1272.5M	2365	1182.6M	2396	1198.0M
1M	1041	1041.0M	1378	1378.8M	1294	1294.5M
2M	626	1252.7M	608	1217.3M	623	1247.3M
4M	315	1260.0M	331	1326.0M	305	1220.3M

4K 100% Random Writes

Silos vs. Solution





PURESTORAGE	XtremIO	IBM	Violin MEMORY	SOLIDFIRE	NIMBUS DATA	skyera	kaminario
Scale Up	Scale Out	Appliance	Appliance	Scale Out	Appliance/Scale Out	Appliance	Scale Out

CISCO

UCS B/C & Invicta

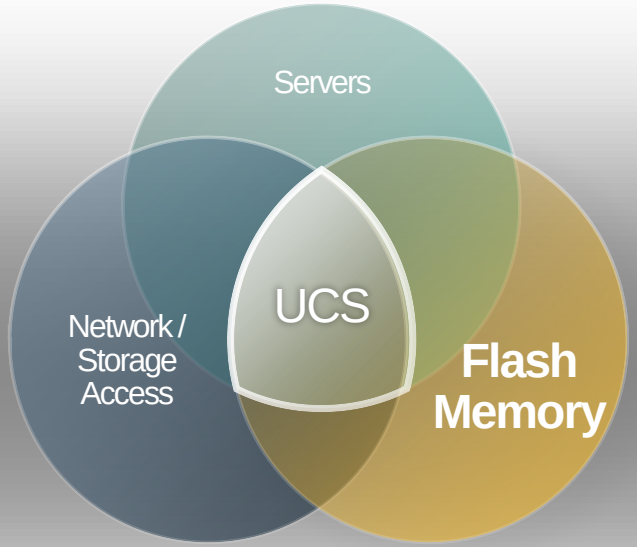
UCS Introduces Flash Memory

UCS Founding Principles

Application Centricity	Operational Simplicity	Platform for IT Innovation

UCS Invicta Series Solid-State Systems

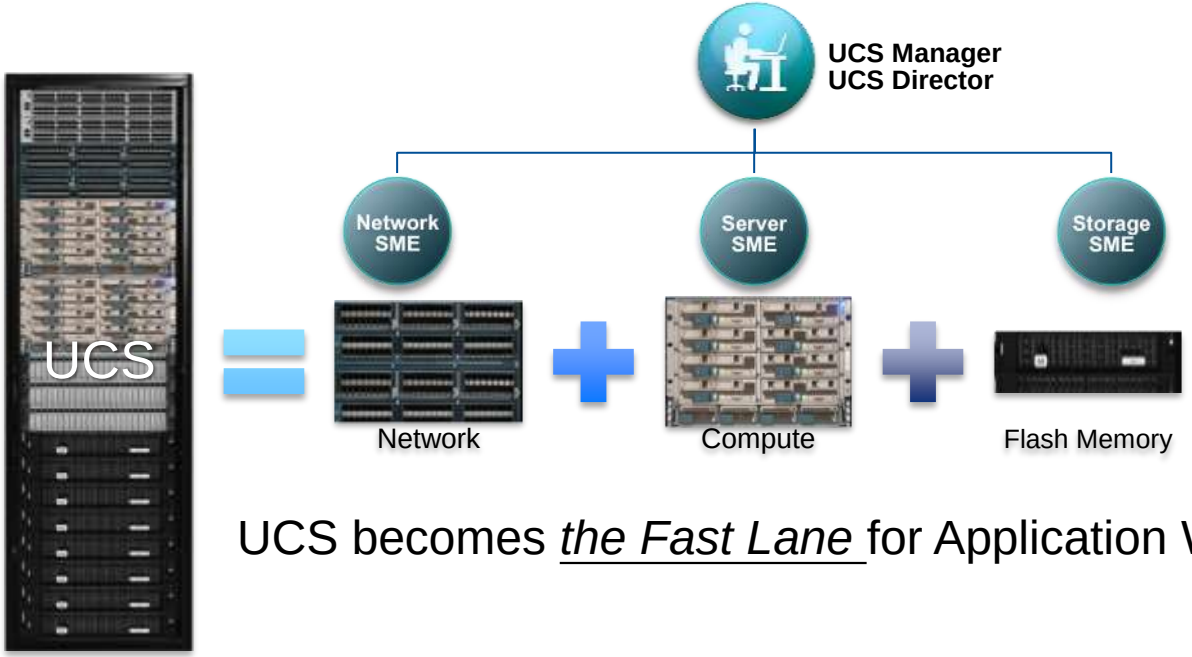
Address new data velocity and scale requirements	Integrate application acceleration into the computing domain



NEXT GEN UNIFIED COMPUTING

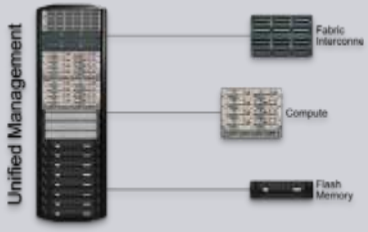
Integration of Solid-State Memory Systems into the UCS Fabric

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

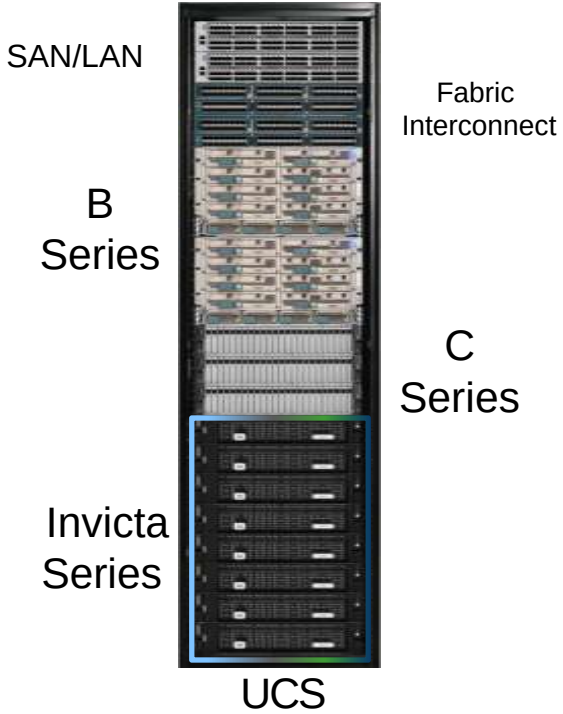
Management

UCS Director

UCS Manager

- Orchestration
- Policies
- Service Profiles
- Self-service
- Tasks
- Workflows

Items in green are completed or in progress



CISCO™