TOMORROW starts here.

Cisco Connect
Praha, Česká republika
24.–25. 3. 2015
Aplikace v kontextu infrastruktury
TECH-DC-2

Martin Diviš, CSE
Agenda

- VCE – Simplified DC Infrastructure
- Cisco Validated Designs
- Infrastructure Considerations
  - MS Private Cloud
  - Big Data
MS Private Cloud
Cloud Infrastructure Requirements

New World – Virtual

Dynamic Pools Of Compute & Storage

- Low-Cost Operations
- Infrastructure-As-A-Service
- Self-Service Consumption Model
- Elastic Resource Allocation
- Shared Pools Of Resources
Microsoft Cloud OS

Public Cloud
- Microsoft Azure
- Azure Virtual Machines
- Windows Azure Pack
- Office 365

Private Cloud
- Service Providers
- Windows Server
- System Center
- SQL Server

Consistent Platform

Source: Microsoft
Microsoft System Center 2012 SP1

Endpoint Protection
Operations Manager
Orchestrator
Service Manager
Configuration Manager
Virtual Machine Manager
App Controller
Data Protection Manager

Source: Microsoft
System Center 2012
Three Broad Capabilities

- Application Management
  - Application Controller
  - VMM
  - Operations Manager

- Service Delivery & Automation
  - Orchestrator
  - Service Manager

- Infrastructure Management
  - VMM
  - Operations Manager
  - DataProtection Mgr
  - Configuration Mgr
Cisco Solutions for MS Private Cloud
Cisco Unified Computing System
Overview

Cisco UCS: Single Object to Manage a Server's Identity

- NIC MACs
- HBA WWNs
- Server UUID
- VLAN Assignments
- VLAN Tagging
- FC Fabric Assignments
- FC Boot Parameters
- Number of vNICs
- Boot order
- PXE settings
- IPMI Settings
- Number of virSHAs
- QoS
- Call Home
- Template Association
- Org & Sub Org Assoc
- Server Role Association
- Static Thresholds
- BIOS scrub actions
- Disk scrub actions
- BIOS firmware
- Adapter firmware
- BMC firmware
- RAID settings
- Advanced NIC settings
- Serial over LAN settings
Hardware “State” Abstraction

- Separate firmware, addresses, and parameter settings from server hardware
- Physical servers become interchangeable hardware components
- Easy to move OS & applications across server hardware
System Center Key Components for UCS

- Application Management
- Service Delivery & Automation
- Infrastructure Management
- Orchestrator
- VMM
- Operations Manager
UCS State View in SCOM
Automated UCS Management with Orchestrator

- Automate UCS management
- Deliver Scalable and Reliable UCS management through Orchestrated Workflows
- Optimize and extend UCS capabilities
UCS Networking for HyperV
UCS Virtual Interface Card

- Up to 256 PCIe devices
  - Devices can be vNICs or vHBAs
  - Each device has a corresponding switch interface

- Bandwidth 2x4x10 Gb
  - Uses 4x10 Ether Channel, HW 40Gb Capable
  - vNICs/vHBAs NOT limited to 10Gb

- PCIe Gen-2 x 16

- Mezzanine and PCIe
Network Configurations
Three options

Non-converged

Converged Option 1

Converged Option 2

Source: Microsoft
# HyperV Interfaces for non-converged

<table>
<thead>
<tr>
<th>Interface Name</th>
<th>FabricFailover</th>
<th>MTU</th>
<th>Fabric</th>
<th>FabricFailover</th>
</tr>
</thead>
<tbody>
<tr>
<td>Management</td>
<td>True</td>
<td>1500</td>
<td>A</td>
<td>Yes</td>
</tr>
<tr>
<td>Livemigration-A</td>
<td>False</td>
<td>9000</td>
<td>A</td>
<td>No</td>
</tr>
<tr>
<td>Livemigration-B</td>
<td>False</td>
<td>9000</td>
<td>B</td>
<td>No</td>
</tr>
<tr>
<td>CSV</td>
<td>False</td>
<td>9000</td>
<td>B</td>
<td>Yes</td>
</tr>
<tr>
<td>VMAccess-A</td>
<td>False</td>
<td>1500</td>
<td>A</td>
<td>No</td>
</tr>
<tr>
<td>VMAccess-B</td>
<td>False</td>
<td>1500</td>
<td>B</td>
<td>No</td>
</tr>
<tr>
<td>SMB3-A</td>
<td>False</td>
<td>9000</td>
<td>A</td>
<td>No</td>
</tr>
<tr>
<td>SMB3-B</td>
<td>False</td>
<td>9000</td>
<td>B</td>
<td>No</td>
</tr>
<tr>
<td>iSCSI-A</td>
<td>False</td>
<td>9000</td>
<td>A</td>
<td>No</td>
</tr>
<tr>
<td>iSCSI-B</td>
<td>False</td>
<td>9000</td>
<td>B</td>
<td>No</td>
</tr>
</tbody>
</table>
### MTU Settings

<table>
<thead>
<tr>
<th>Priority</th>
<th>Enabled</th>
<th>CoS</th>
<th>Packet Drop</th>
<th>Weight</th>
<th>Weight (%)</th>
<th>MTU</th>
<th>Multicast Optimized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Platinum</td>
<td>✔️</td>
<td>5</td>
<td></td>
<td>10</td>
<td>N/A</td>
<td>9000</td>
<td></td>
</tr>
<tr>
<td>Gold</td>
<td></td>
<td>4</td>
<td>✔️</td>
<td>9</td>
<td>N/A</td>
<td>normal</td>
<td></td>
</tr>
<tr>
<td>Silver</td>
<td></td>
<td>2</td>
<td>✔️</td>
<td>8</td>
<td>N/A</td>
<td>normal</td>
<td></td>
</tr>
<tr>
<td>Bronze</td>
<td></td>
<td>1</td>
<td>✔️</td>
<td>7</td>
<td>N/A</td>
<td>normal</td>
<td></td>
</tr>
<tr>
<td>Best Effort</td>
<td>✔️</td>
<td>Any</td>
<td>✔️</td>
<td>5</td>
<td>50</td>
<td>normal</td>
<td></td>
</tr>
<tr>
<td>Fibre Channel</td>
<td>✔️</td>
<td>3</td>
<td></td>
<td>5</td>
<td>50</td>
<td>fc</td>
<td>N/A</td>
</tr>
</tbody>
</table>
MTU Settings

Create QoS Policy

Name: Hyper-V

Priority: Platinum
MTU Settings

Create QoS Policy

Create VLAN

MTU: [9000]

Warning

Make sure that the MTU has the same value in the QoS System Class corresponding to the Egress priority of the selected QoS Policy.

MAC Pool: <not set>

QoS Policy: Hyper-V
Advanced Networking Features
Virtual Machine Queue

Default

VM Queue & Receiver Side Scaling Enabled
VMQ Configuration

LAN Cloud -> VMQ Connection Policies

Number of Interrupts = CPU Cores * CPUs * 2
VMQ: Max number 128
BIOS Policy: SRIOV
Adapter Policy: Windows
NVGRE Offloading

Stateless offloads with NVGRE cannot be used with NetFlow, usNIC, VM-FEX, or VMQ.

Only supported on VIC1340 and VIC1380 with MS Server 2012 R2.

Transmit Queues = 1
Receive Queues = n (up to 8)
Completion Queues = #Transmit Q + #Receive Q
Interrupts = #Completion Q + 2

Network Virtualization using Generic Routing Encapsulation = Enabled
Interrupt Mode = Msi-X
Virtual Networking for MS Private Cloud
Hyper-V Extensible Switch

Source: Microsoft
Cisco Virtual Networking Solutions

Bring network to the hypervisor (Cisco Nexus 1000V Switch)

Bring VM awareness to physical network (Cisco UCS VM-FEX)
Virtual Networking with SR-IOV
UCS VM-FEX Modes of Operation

- Windows Hyper-V with SR-IOV

### Emulated Mode
- Each VM gets a dedicated PCIe device
- Appears as distributed virtual switch to hypervisor
- LiveMigration supported

### Hypervisor Bypass
- Co-exists with Standard mode
- Bypasses Hypervisor layer
- ~30% improvement in I/O performance
- Appears as distributed virtual switch to hypervisor
- Currently supported through SR-IOV with Hyper-V 2012 Live Migration supported
Lower latency and end results

- Test: TPC-H database with 100 Scale factor
- VM configuration: 16 vCPU, 32G RAM
- SQL server 2012: TPC-H database on 2 LUNs
- Metric: IO response time (under 20ms for read IO), Query response time

![vSwitch vs. VM-FEX - Disk Access Latency (in sec)](chart)

[www.cisco.com/go/vmfex](http://www.cisco.com/go/vmfex)
Big Data – Infrastructure Considerations
“More data usually beats better algorithms.”

-Anand Rajaraman, SVP @WalmartLabs
Traditional Enterprise Data Management

- Operational (OLTP)
- Operational (OLTP)
- Operational (OLTP)

Online Transactional Processing

Extract, Transform, and Load (batch processing)

Enterprise Data Warehouse

Business Intelligence

ETL
So what has changed?

- The Explosion of Unstructured Data

- **1.8 trillion gigabytes** of data was created in 2011...
  - More than 90% is unstructured data
  - Approx. 500 quadrillion files
  - Quantity doubles every 2 years
  - Most unstructured data is **neither stored nor analyzed**!

Source: Cloudera
Enterprise Data Management with Big Data

- Operational (OLTP)
- Operational (OLTP)
- Operational (OLTP)
- Web
- Machine
- Big Data (Hadoop, etc.)
- ETL
- EDW
- BI/Reports
- Dashboards
What is Big Data?

When the size of the data itself is part of the problem.
Classic NAS/SAN vs. new scale-out DAS

Traditional – separate compute from storage

New – move the compute to the storage

Low-cost, DAS-based, scale-out clustered filesystem
Big Data Software Architectures and Design Considerations
Three Common Big Data Architectures

NoSQL
Fast key-value store/retrieve in real time

Hadoop
Distributed batch, query, and processing platform

MPP Relational Database
Scale-out BI/DW
Hadoop is a distributed, fault-tolerant framework for storing and analyzing data. Its two primary components are the Hadoop Filesystem (HDFS) and the MapReduce application engine.
Hadoop: A Closer Look

- Hadoop 2.0 (with YARN) adds the ability to run additional distributed application engines concurrently on the same underlying filesystem.
MapReduce Example: Word Count Frequency

- **Input**
  - the quick brown fox
  - the fox ate the mouse
  - how now brown cow

- **Map**
  - the, 1
  - brown, 1
  - fox, 1
  - quick, 1

- **Shuffle & Sort**
  - the, 1
  - brown, 1
  - fox, 1
  - how, 1
  - now, 1
  - quick, 1

- **Reduce**
  - brown, 2
  - fox, 2
  - how, 1
  - now, 1
  - the, 3
  - ate, 1
  - cow, 1
  - mouse, 1
  - quick, 1

- **Output**
  - Hadoop

- **Command Line Tools**
  - cat
  - grep
  - sort
  - uniq
Hadoop Components and Operations

Hadoop Distributed File System
- Scalable & Fault Tolerant
- Filesystem is distributed, stored across all data nodes in the cluster
- Files are divided into multiple large blocks – 64MB default, typically 128MB – 512MB
- Data is stored reliably. Each block is replicated 3 times by default
- Types of Nodes
  - Name Node - Manages HDFS
  - Job Tracker – Manages MapReduce Jobs
  - Data Node/Task Tracker – stores blocks/does work
HDFS Architecture

Switch

Name Node

/usr/sean/foo.txt:blk_1,blk_2
/usr/jacob/bar.txt:blk_3,blk_4

Data node 1: blk_1
Data node 2: blk_2, blk_3
Data node 3: blk_3
Data node 4: blk_1, blk_2, blk_3
Data node 5: blk_4
Hadoop and the Network

- Job types and their network traffic patterns
- Network characteristics of the cluster
- Impact of QoS
Hadoop Network Design

- The network is the fabric – the ‘bus’ - of the ‘supercomputer’

- Big data clusters often create **high east-west, any-to-any** traffic flows compared to traditional DC networks

- Hadoop networks are typically isolated/dedicated; simple leaf-spine designs are ideal

- 10GE typical from server to ToR, low oversubscription from ToR to spine

- With Hadoop 2.0, clusters will likely have heterogeneous, multi-workload behavior
Hadoop network traffic types

**Small Flows/Messaging**
(Admin Related, Heart-beats, Keep-alive, delay sensitive application messaging)

**Small – Medium Incast**
(Hadoop Shuffle)

**Large Flows**
(HDFS egress)

**Large Pipeline**
(Hadoop Replication)
Map and Reduce Traffic

Many-to-Many Traffic Pattern

Map 1 — Map 2 — Map 3 — Map N
Reducer 1 — Reducer 2 — Reducer 3 — Reducer N
HDFS
Analyze Workload
Wordcount on 200K Copies of complete works of Shakespeare

Note:
Due the combination of the length of the Map phase and the reduced data set being shuffled, the network is being utilized throughout the job, but by a limited amount.

The red line is the total amount of traffic received by hpc064

These symbols represent a node sending traffic to HPC064

Network graph of all traffic received on a single node (80 node run)
Transform Workload (1TB Terasort)

Network graph of all traffic received on a single node (80 node run)

These symbols represent a node sending traffic to HPC064.

The red line is the total amount of traffic received by hpc064.
Output Data Replication Enabled

- Replication of 3 enabled (1 copy stored locally, 2 stored remotely)
- Each reduce output is replicated now, instead of just stored locally

Note:
If output replication is enabled, then at the end of the job HDFS must store additional copies. For a 1TB sort, additional 2TB will need to be replicated across the network.
Job Patterns - summary

Job Patterns have varying impact on network utilization

- **Analyze**
  Simulated with Shakespeare Wordcount

- **Extract Transform Load (ETL)**
  Simulated with Yahoo TeraSort

- **Extract Transform Load (ETL)**
  Simulated with Yahoo TeraSort with output replication
Frequently Asked Questions
Does Hadoop really need 10GE?

Definitely, so tune for it!

- Analytic workloads tend to be lighter on the network
- Transform workloads tend to be heavier on the network
- Hadoop has numerous parameters which affect network
- Take advantage of 10GE:
  - mapred.reduce.slowstart.completed.maps
  - dfs.balance.bandwidthPerSec
  - mapred.reduce.parallel.copies
  - mapred.reduce.tasks
  - mapred.tasktracker.reduce.tasks.maximum
  - mapred.compress.map.output
Consistent, low network latency is desirable, but ultra low latency does not represent a significant factor for typical Hadoop workloads.

Note: There is a difference in network latency vs. application latency. Optimization in the application stack can decrease application latency that can potentially have a significant benefit.
Can QoS help?
An example with HBase
HBase + MapReduce with QoS

~60% Read Improvement

Read Latency
Comparison of Non-QoS vs. QoS Policy

Switch Buffer Usage
With Network QoS Policy to prioritize HBase Update/Read Operations
ACI Fabric Load Balancing
Flowlet Switching

- **Flowlet switching** routes bursts of packets from the same flow independently, based on measured congestion of both external wires and internal ASICs.
- Allows packets from the same flow to take different paths, while maintaining packet ordering.
- Provides better (more evenly distributed) utilization of available paths.
- Does all this transparently – *nothing to modify at the host/app level*.
Real traffic is a mix of large (elephant) and small (mice) flows.

Key Idea: Fabric detects initial few flowlets of each flow and assigns them to a high priority class.

Standard (single priority): Large flows severely impact performance (latency & loss), for small flows

Dynamic Flow Prioritization: Fabric automatically gives a higher priority to small flows.
Network Summary

- The network is the “system bus” of the Hadoop “supercomputer”
- Analytic- and ETL-style workloads can behave very differently on the network
- Ultra-low latency probably not critical for typical Hadoop workloads
- Minimize oversubscription, leverage QoS and DPP, and tune Hadoop to take advantage of 10GE – *distribute fairly*
Cisco UCS and Big Data

- Building a big data cluster with the UCS Common Platform Architecture (CPA)
- CPA Networking
- CPA Sizing and Scaling
Hadoop Server Hardware Evolving in the Enterprise

Typical 2009 Hadoop node
- 1RU server
- 4 x 1TB 3.5” spindles
- 2 x 4-core CPU
- 1 x GE
- 24 GB RAM
- Single PSU
- Running Apache
- $

Economics favor “fat” nodes
- 6x-9x more data/node
- 3x-6x more IOPS/node
- Saturated gigabit, 10GE on the rise
- Fewer total nodes lowers licensing/support costs
- Increased significance of node and switch failure
- $$$

Typical 2015 Hadoop node
- 2RU server
- 12 x 4TB 3.5” or 24 x 1TB 2.5” spindles
- 2 x 6-12 core CPU
- 2 x 10GE
- 128-256 GB RAM
- Dual PSU
- Running commercial/licensed distribution
- $$$$
Frequently Asked Questions
Hadoop and JBOD
Why not use RAID-5?

- It hurts performance:
  - RAID-5 turns parallel sequential reads into slower *random* reads
  - RAID-5 means speed limited to the *slowest* device in the group
- It’s wasteful: Hadoop already replicates data, no need for more replication
  - Hadoop block copies serve **two** purposes: 1) redundancy and 2) performance (more copies available increases data locality % for map tasks)
Can I virtualize?
Yes you can (easy with UCS), but should you?

- Hadoop and most big data architectures can run virtualized
- However this is typically not recommended for performance reasons
  - Virtualized data nodes will contend for storage and network I/O
  - Hypervisor adds overhead, typically without benefit
- Some customers are running master/admin nodes (e.g. Name Node, Job Tracker, Zookeeper, gateways, etc.) in VM’s, but consider single point of failure
- UCS is ideal for virtualization if you go this route
CPA Network Design for Big Data
Cisco UCS Common Platform Architecture (CPA)

Building Blocks for Big Data

- UCS Manager
- Nexus 2232 Fabric Extenders (optional)
- UCS 6200 Series Fabric Interconnects
- LAN, SAN, Management
- UCS C220/C240 M4 Servers
New UCS Reference Configurations for Big Data

**Quarter-Rack UCS Solution for MPP, NoSQL – High Performance**
- 2 x UCS 6248
- 8 x C220 M4 (SFF)
- 2 x E5-2680v3
- 256GB
- 6 x 400-GB SAS SSD

**Full Rack UCS Solution for Hadoop, NoSQL – Balanced**
- 2 x UCS 6296
- 16 x C240 M4 (SFF)
- 2 x E5-2680v3
- 256GB
- 24 x 1.2TB 10K SAS

**Full Rack UCS Solution for Hadoop Capacity-Optimized**
- 2 x UCS 6296
- 16 x C240 M4 (LFF)
- 2 x E5-2620v3
- 128GB
- 12 x 4TB 7.2K SATA
CPA Recommended FEX Connectivity

- 2232 FEX has 4 buffer groups: ports 1-8, 9-16, 17-24, 25-32
- Distribute servers across port groups to maximize buffer performance and predictably distribute static pinning on uplinks
“NIC bonding is one of Cloudera’s highest case drivers for misconfigurations.”

http://blog.cloudera.com/blog/2015/01/how-to-deploy-apache-hadoop-clusters-like-a-boss/
UCS Fabric Failover

- Fabric provides NIC failover capabilities chosen when defining a service profile
- **Avoids traditional NIC bonding in the OS**
- Provides failover for both unicast and multicast traffic
- Works for any OS on bare metal
- (Also works for any hypervisor-based servers)
Recommended UCS networking with Apache Hadoop

- Use 2 VNICS with Fabric Failover on opposite fabrics for internal and external traffic

- VNIC 1 on Fabric A with FF to B (internal cluster)
- VNIC 2 on Fabric B with FF to A (external data)
- No OS bonding required
- VNIC 0 (management) wiring not shown for clarity (primary on Fabric B, FF to A)

Note: cluster traffic will flow northbound in the event of a VNIC1 failover. Ensure appropriate bandwidth/topology.
Summary

- CVD's
- MS Cloud – UCS advantage through service profiles, advanced networking options
- Big Data – Network and Server (CPA) considerations