



# Overview

Ghislain de Jacquilot

March 22, 2010

# Business Model

## Vertical Solutions Approach



## Premier Server OEM Go-To-Market Strategy



## Differentiated Products





# Hardware

March 22, 2010

# Grid Director 4000 Series



4036



4036E



4200

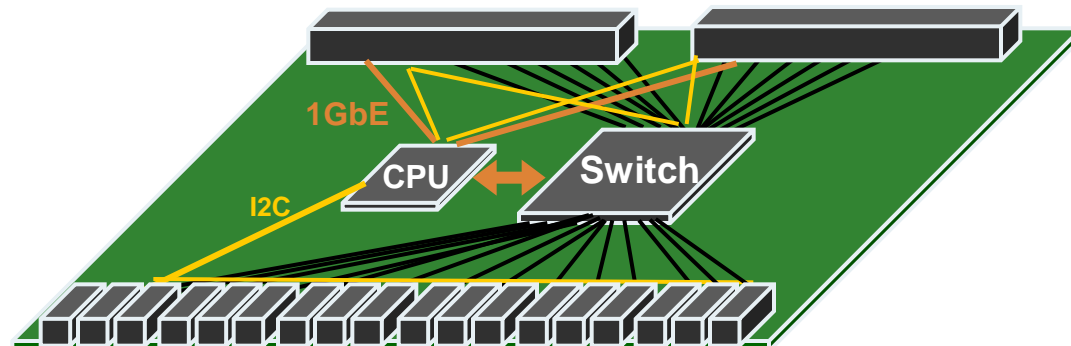


4700

- ▶ **Scalable** QDR switch family, with a **modular building block** approach
- ▶ **Extreme low latency**
- ▶ **Smart managed** switches with **advanced fabric management**
- ▶ Designed for **commercial-grade** Reliability, Availability and Manageability
- ▶ Most **mature**, 4<sup>th</sup> Generation switch family and switch silicon

## Line & Fabric board Architecture

- ▶ **Unique Line/Fabric Board Architecture including on board CPU**
  - Supports Voltaire’s robust cable/signal optimization mechanism
    - Automatically configure SERDES parameters on the fly, as new cables are inserted
    - Allows maximum support for different cable types/vendors/lengths/gauges
  - Ready for supporting congestion management and dynamic routing
    - Monitor congestion, traffic counters, analyze and export statistics to UFM
    - React immediately to local routing failures and changes
    - Maximize scalability by offloading heavy processing from the management board CPU
    - Assist for MPI collective offload operations
- ▶ **Redundant Out of Band 1Gb/s Ethernet for Management interface**
- ▶ **Redundant I2C Interface for low level indications**



# CPU, Memory, Storage

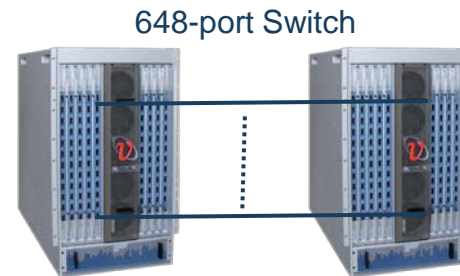
- ▶ **AMCC PowerPC 460EX embedded CPU**
  - PCI-E interfaces (connected to the Shaldag device)
- ▶ **512MB of DRAM with ECC support (SORDIMM)**
- ▶ **64MB NOR Flash**
  - Boot
  - Linux Kernel
  - Primary file system (switch SW)
  - Secondary file system (redundant switch SW)
- ▶ **256MB NAND flash**
  - Log files

# Voltaire Grid Director 4700

## HyperScale™ Unique Stackable Architecture



### HyperScale-Twin



- ▶ 108 CXP-CXP cables
- ▶ Fully non-blocking
- ▶ Flexible Racking

- ▶ Replace standard fabric boards with HyperScale™ fabric boards
- ▶ HyperScale™ fabric boards contain double the capacity as well as external 12X CXP ports



# Voltaire Tree-Mesh Architecture

## Example 2,268 Nodes Configuration (can expand to any size)

### ▶ Lowest Latency

- 400ns Vs. 750ns with competition

### ▶ Simpler Cabling

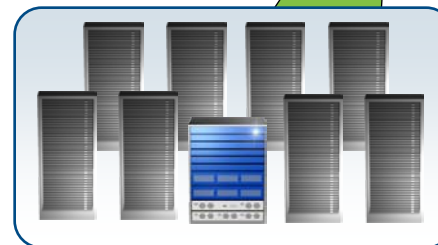
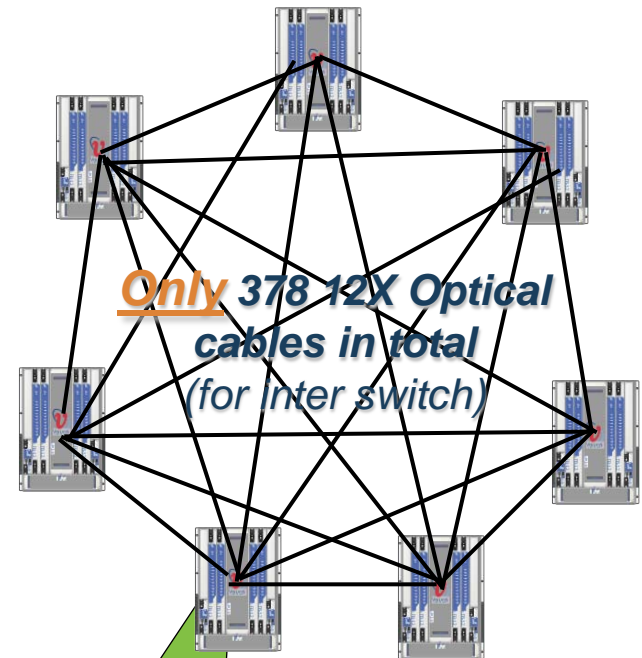
- 1/6 of the cables needed between switches
- No concentration at spine level

### ▶ 100-50% bisectional BW at any scale

- 100% bandwidth per ~7800 core island (Most jobs don't scale that much)
- Min of 50% bandwidth between island or 75% if considering PCIe limitations , Can get to 100% with less line cards

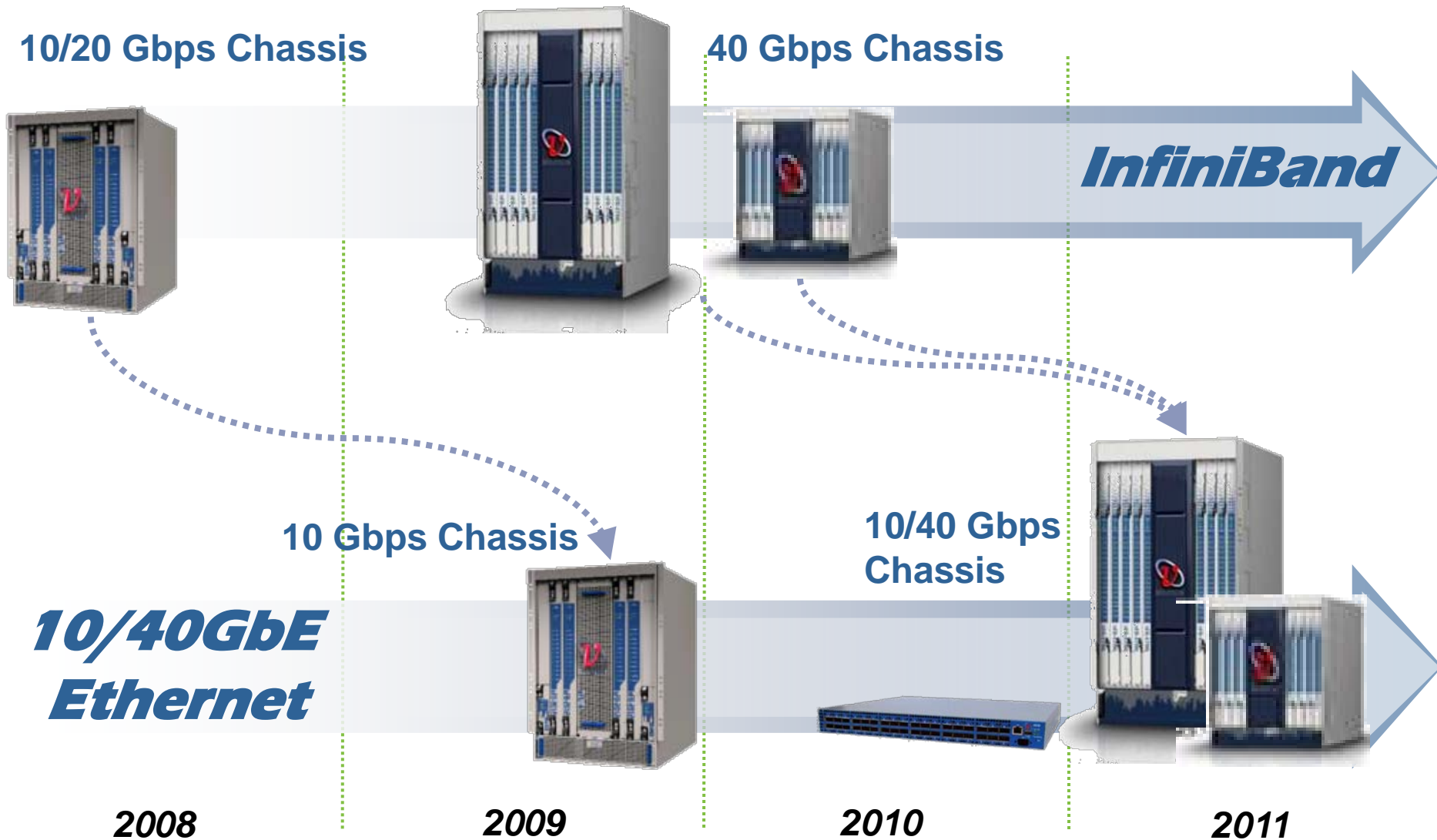
### ▶ Integrated with Voltaire UFM

*2,268 nodes built out of 7 islands*

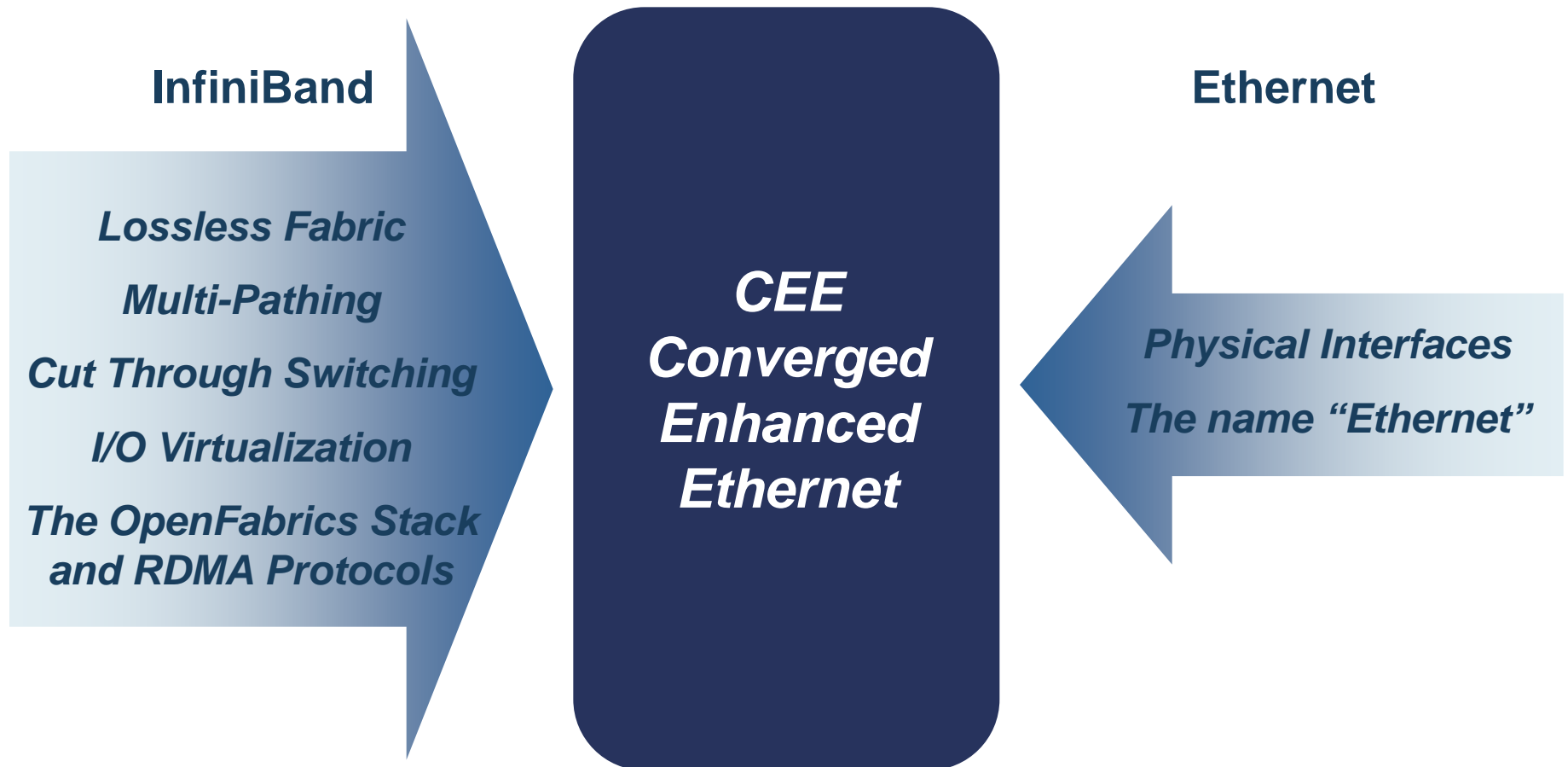


*Individual Island of 7776 CPUs (24 cores \* 324)*

# Leveraging Voltaire's InfiniBand Expertise



# Technology Convergence: The New Ethernet



***InfiniBand Expertise is a Key Advantage in CEE***

# CEE/DCE/DCB

## InfiniBand-like Features under the Ethernet Name

Standard	Benefit
<b>802.1Qbb</b> <b>Priority Based Flow Control</b>	<b>Provides Class of Service per flow (type of traffic)</b>
<b>802.1Qaz</b> <b>Enhanced Transmission Selection</b>	<b>Allows lower priority traffic to use unused bandwidth from the high-priority queues</b>
<b>802.1Qau</b> <b>Congestion Notification</b>	<b>Allows congested points to request that ingress ports limit their transmission when congestion is occurring</b>
<b>DCBX</b>	<b>Ensures consistent configuration across the network</b>
<b>L2 Multipathing</b>	<b>Eliminate STP for L2 topologies, utilizing full bi-sectional bandwidth, faster failover</b>

# Introducing Voltaire Vantage™ 8500

## ▶ Highest Capacity Chassis

- Leveraging Voltaire 11.5Tb/s chassis
- Total 10GbE-port options: 288 (full rate)
- Flat Layer 2 Scalability to 3,400 nodes (full rate)

## ▶ Key Features

- Low latency (<1us) and low power (<10W/Port)
- Ethernet L2 Multi-pathing
- Virtual I/O port configuration per Server VM
- Full L2 stack, L3 ready, L2-4 ACLs
- Enhanced Ethernet (CEE) feature set
- Advanced congestion control and monitoring





# Software

March 22, 2010

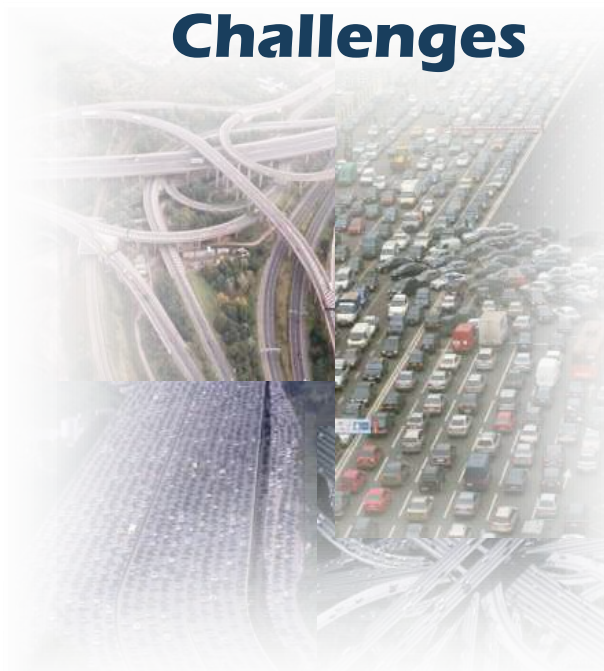
# Data Center Challenges

## Complexity



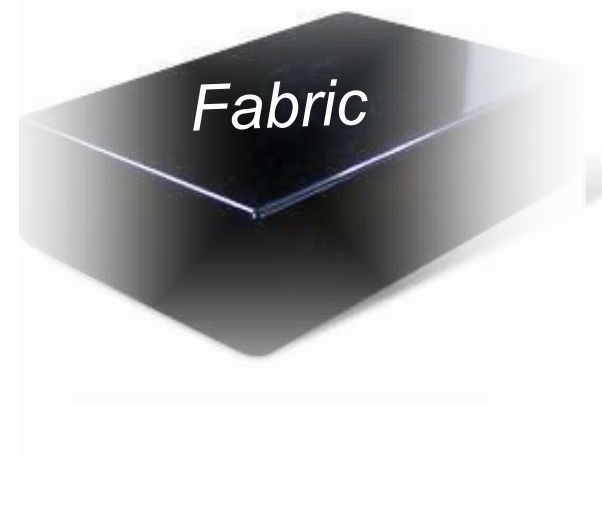
- Traditional device management is not enough

## Traffic Challenges



- Benchmarks differ from real-life traffic patterns

## Lack of Visibility



- Undetected issues, unutilized fabric, OPEX burden

# An Infiniband Fabric is not a black box (1/2)

## ► Requires Hardware management

- Detect failures, communication problems
  - Inside the Infiniband Fabric
    - Port counters
    - Port status (QDR,DDR,SDR – 4X,2X,1X)
    - Firmware upgrades (Switch and HCA ASICs)
  - Outside the Infiniband Fabric
    - Chassis
    - Power supplies
    - Fans
    - Temperature
    - Chassis software updates (Switch management)

# An Infiniband Fabric is not a black box (2/2)

- ▶ **What about performance ?**
- ▶ **Some embarrassing questions...**
  - Blocking vs non-blocking fabrics ?
  - Influence of routing algorithms ?
  - Congestion ?
  - Mixing different protocols on the same fabric ?
  - Running multiple jobs on the same fabric ?
  - Performance monitoring Tools ?

# UFM Central Management Platform

## ▶ In-depth visibility into fabric health and traffic

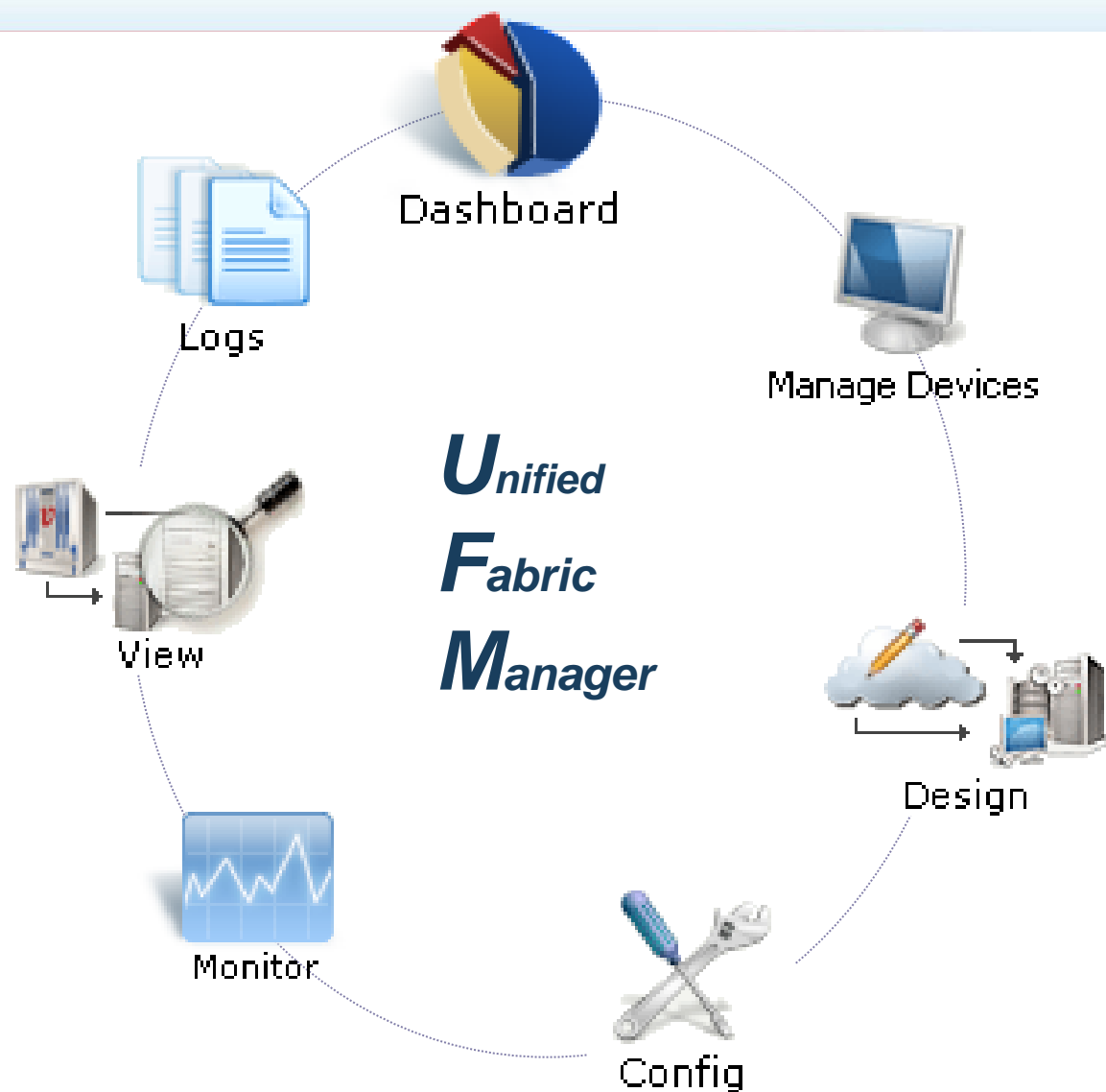
- Central Dashboard, Unique Congestion Map
- Advanced monitoring engine, threshold based alerts

## ▶ Optimize application performance

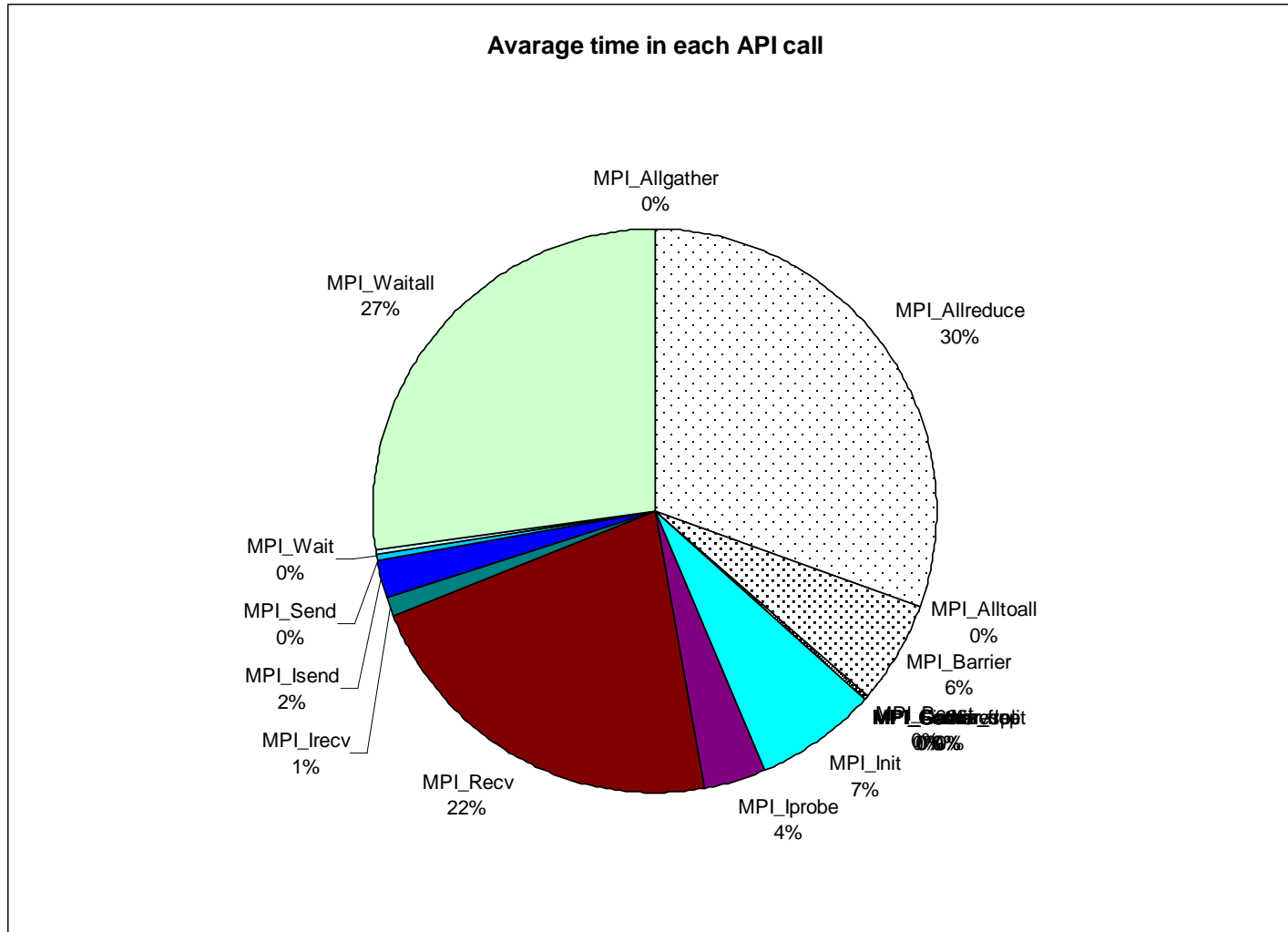
- Quality of Service
- Traffic Aware Routing Algorithm

## ▶ Efficient operations of thousands of fabric components

- Automated configuration of hosts and switches, group tasks
- Seamless change management



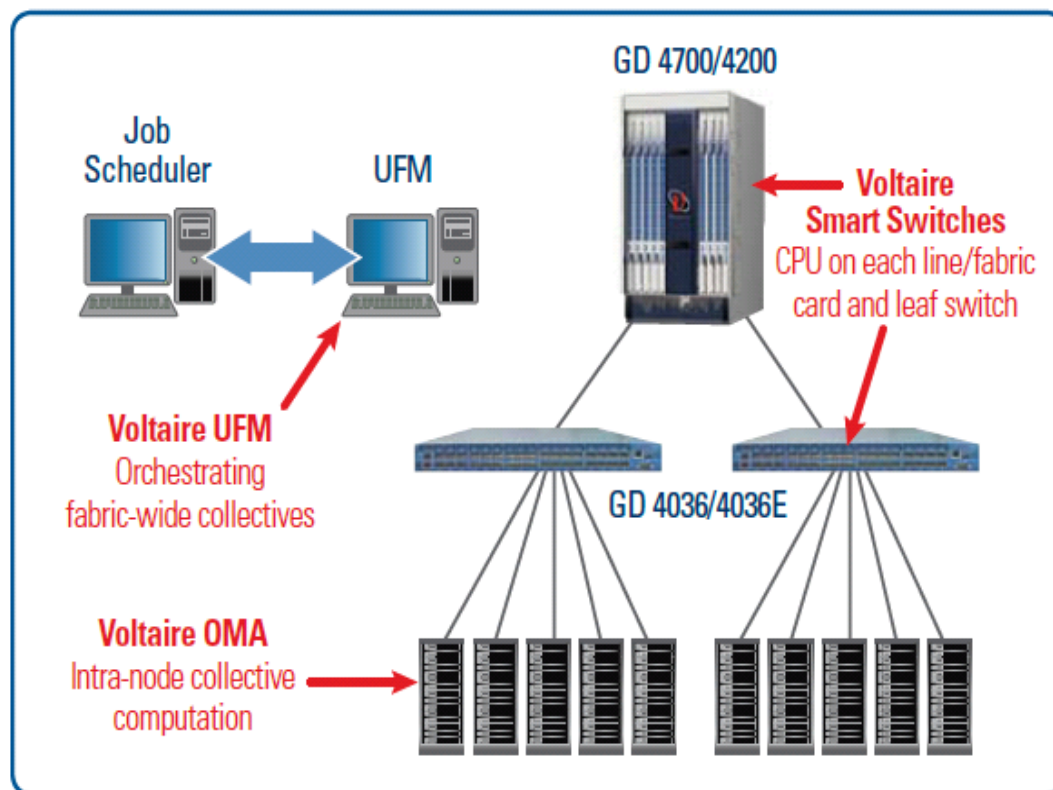
# Fluent (eddy\_417k): 32 Ranks MPI API calls, % of time ranks spent in each call



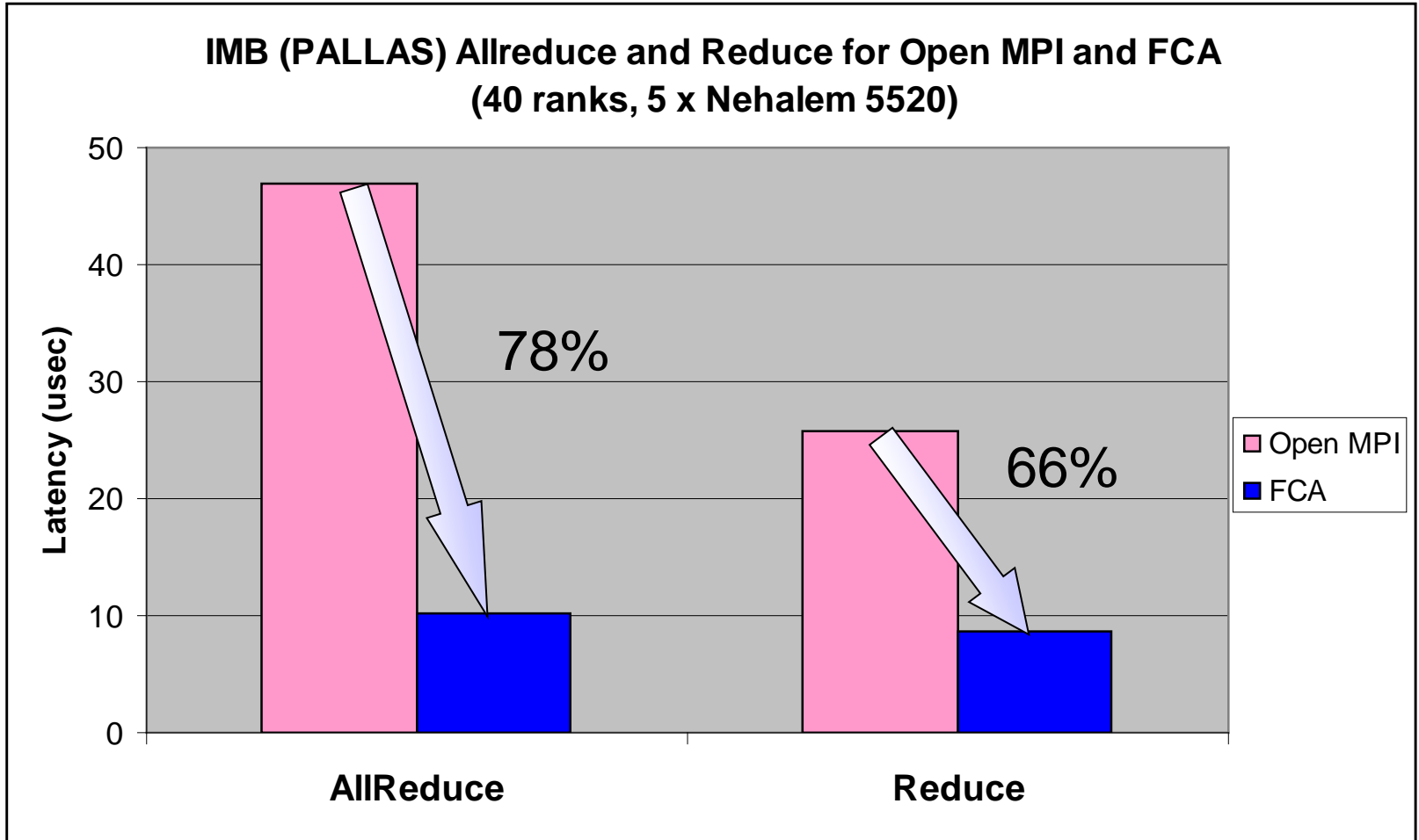
# What is Voltaire FCA ?

## First Fully Integrated solution to offload collectives combine intelligence on server, switches, and management

- UFM™ Automate fabric collective offload/monitoring and integrate with schedulers
- Voltaire “smart” switch based CPUs performing reduction and messaging operation
- Voltaire OMA (Open MPI plug-in) address server side



# FCA Preliminary Performance Results



# FCA results on 512 ranks

	OpenMPI	Voltaire FCA	% improve
Allreduce	145	22	85%
Reduce	57	21	64%
Barrier	139	22	84%

