



H L R I S 

NEC



# Achieving Peak Performance with Advanced Fabric Management *A Case Study with HLRS and NEC*

June 21, 2010

# In this session

## ▶ Tag team presentation by

- HLRS, Stuttgart – The End User - **Uwe Küster**
- NEC – HPC System Provider and Operator - **Dr. Andreas Findling**
- Voltaire – Scale-out Fabric Solutions Provider – **Yaron Haviv**

H L R I S 

**NEC**

VOLTAIRE

## ▶ About the team

## ▶ Large scale systems scalability challenges

## ▶ How connecting schedulers and the fabric manager can improve performance, scalability and overall efficiency

## ▶ Future directions

# Experiences with advanced Voltaire InfiniBand components at HLRS

Uwe Küster(HLRS), Holger Berger(NEC), Bernd Krischok(HLRS)

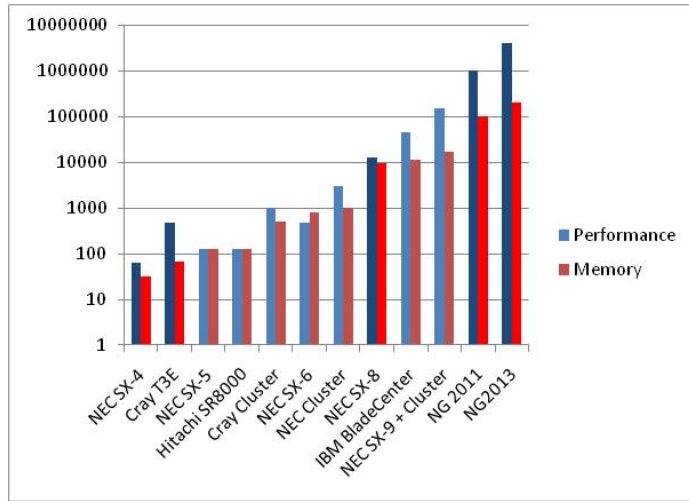


## The Role of HLRS

- Central Unit of the University of Stuttgart
  - Supercomputing since 1962
  - First Cray System in 1982
- 1<sup>st</sup> German National Supercomputing Center
  - Founded 1996
  - Partner of Jülich and Munich
  - Foundation of Gauss Center for Supercomputing 2007



# The HLRS Systems



IBM HSM  
> 2 PB  
Fileserver



NEC SX-9 + Cluster  
(176+1400\*4)



NEC Asama (64)



SUN Fire 2900 (144x2)  
(48x2)

IBM x3755



Cray XT5+XD1  
(224\*4+48)



DALCO Viz-Cluster  
Opteron (64)

IBM  
Cell



IBM  
(2800x4)

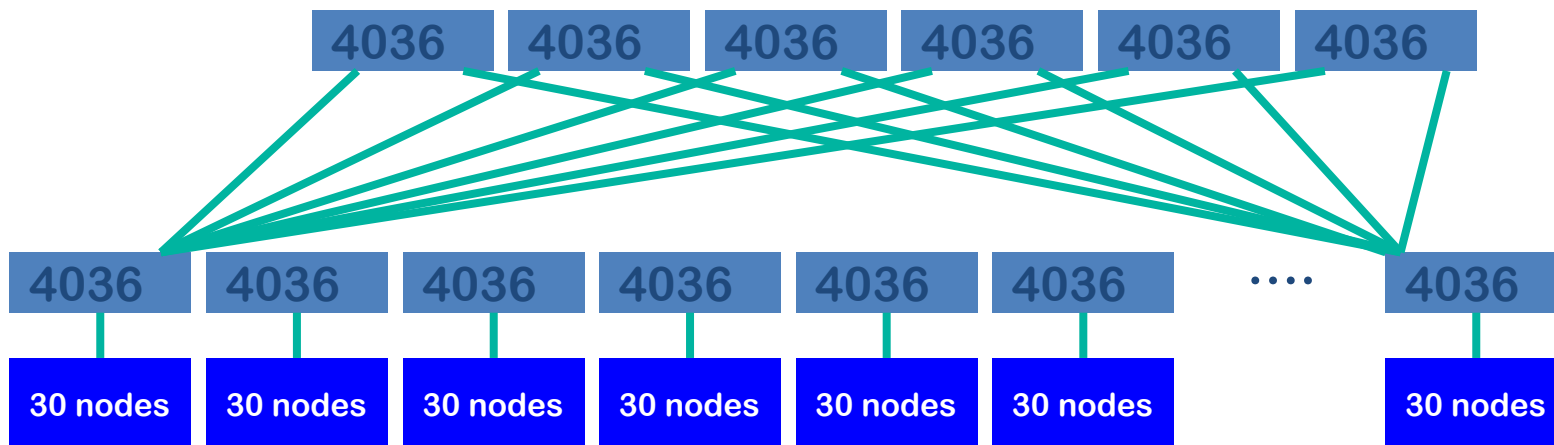


## HLRS experiences with Voltaire Infiniband

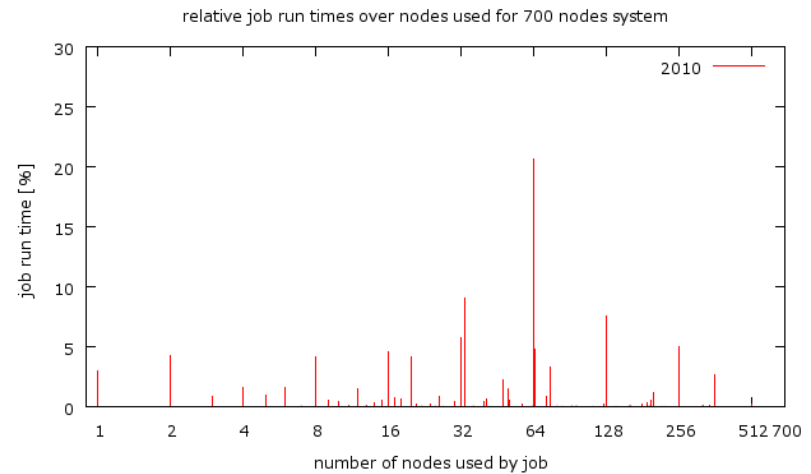
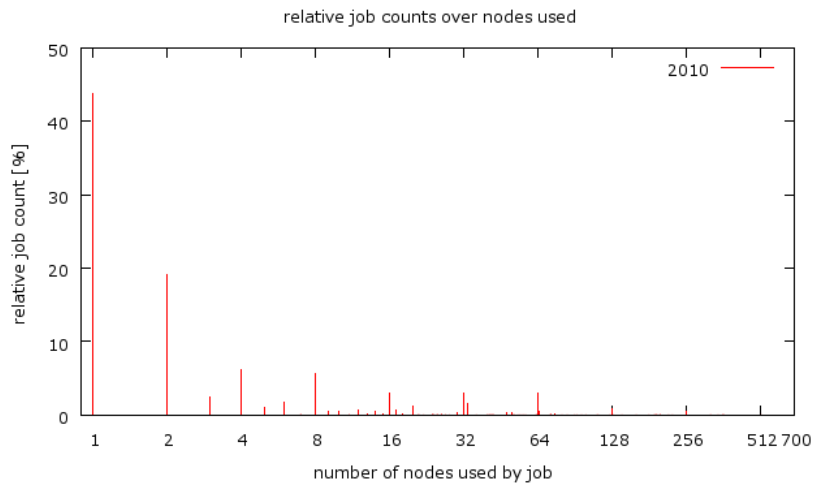
- Now for 5 years in HLRS
  - Cluster Cacau with Voltaire SDR switch (2005)
  - Cluster BW-Grid with Voltaire DDR switches (2008)
  - Cluster Nehalem (Prace Prototype) with Voltaire QDR switches (2009)
- Long time experience in daily production
- Stable operations
- Better signals → longer cables
- „Open Subnet Manager“ essentially programmed by Voltaire
- Latest innovation in software products
  - Open MPI Accelerator (OMA)
  - Unified Fabric Manager (UFM)
  - Fabric Collective Accelerator (FCA)

## NEC LX cluster at HLRS (3)

- Installed 1 year ago, PRACE Prototype hybrid machine combined with NEC SX-9
- 700 nodes with 2.8 Ghz Nehalem CPUs, memory/node: 12, 24, 48, 128, 144 GB
- 32 nodes with 32 Tesla S1070 → 64 GPGPUs
- 24 leaf + 6 spine switches Voltaire 4036 QDR Switches (36 ports)
- Nodes with Infinihost III *DDR* HCAs
- Voltaire UFM Unified Fabric Manager



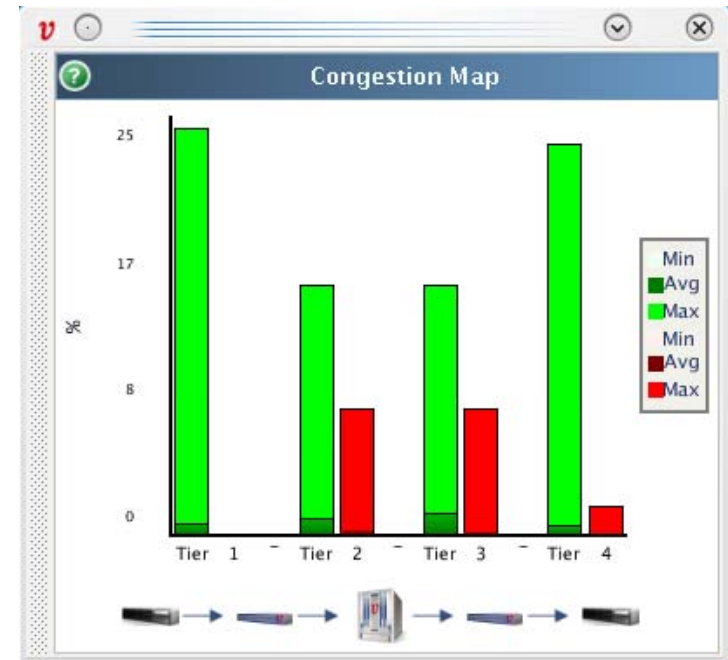
# Parallel load distribution on HLRS Nehalem cluster



- Most jobs are single node jobs (= 8 cores)
- Most run time is consumed by 64 nodes jobs (=512 cores)
- Sweet spot:
  - Small job wait time for user
  - No need to for discharging the machine to run large jobs
  - Most jobs are long running jobs ( → difficult backfilling)

## Inexpensive Blocking Network

- 30 DDR versus 6 QDR uplinks  
→ blocking factor 5:2
- Solution was not optimal but cost effective
  - Small number of small relative switches
  - Small number of long cables
- Is that sufficient? How large is
  - congestion?
  - the load of the spine switches?
  - the burst load?
  - the average load?
- Optimization is achievable with right tools
  - The Voltaire UFM congestion map indicates a reasonable behaviour for average congestion and average load



# NEC LX-Series Supercomputer



# NEC's HPC Offerings - 2010

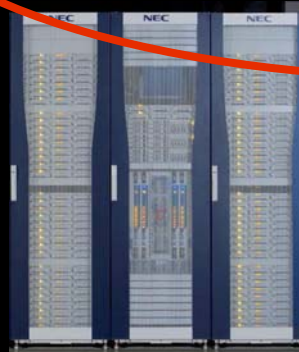
VECTOR

SX-9



x86

LX-Series

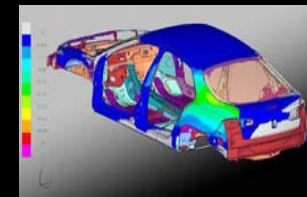


HPC Filesystem

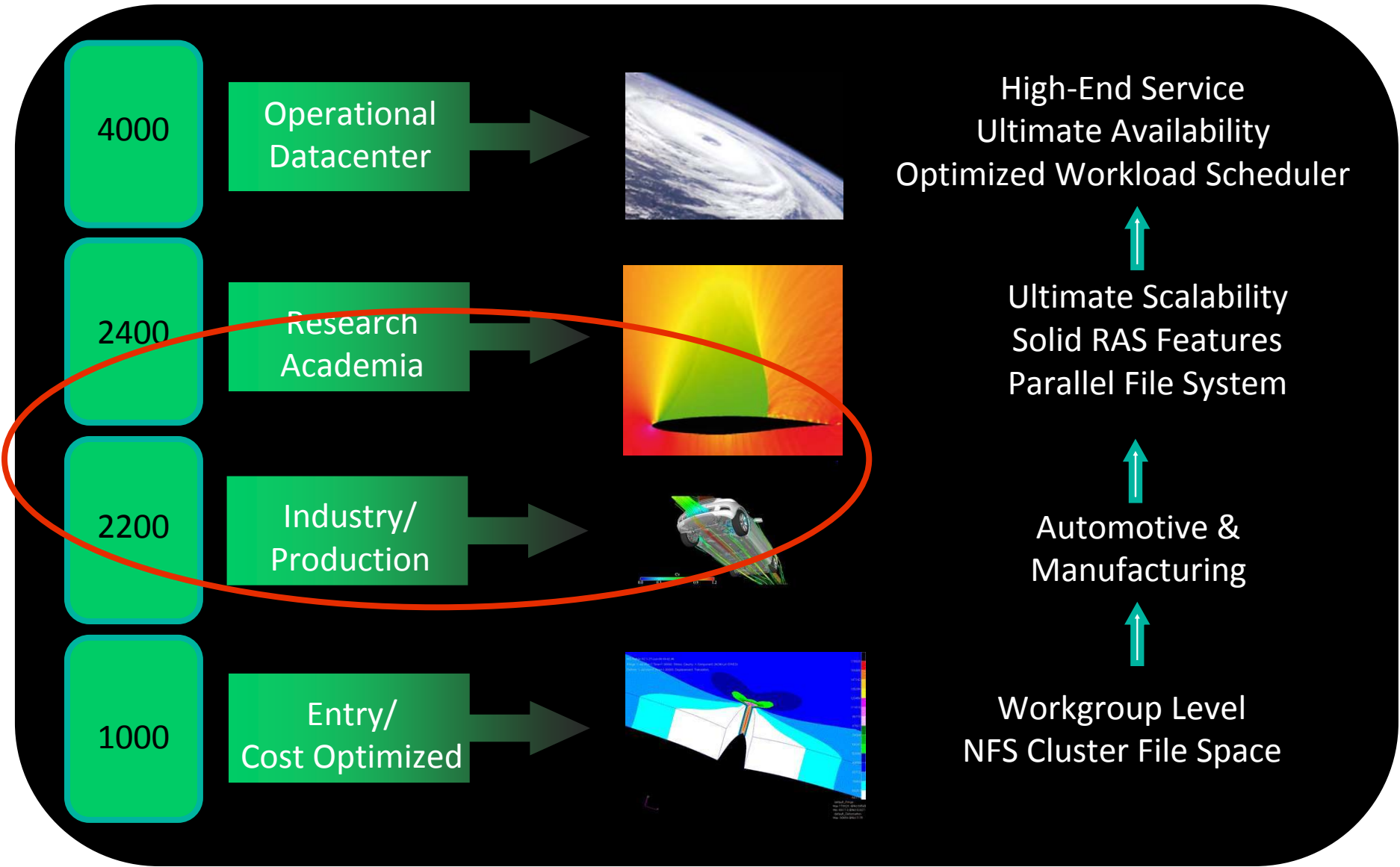
LXFS



Professional Services



# NEC LX Product Options



# NEC LX 2400: Rank 20 on Green500

Installed at HLRS in June 2009  
700 nodes, Nehalem-EP  
Infiniband Fabric, LXFS  
62 Tflops Peak  
+ 32 NVIDIA Tesla S1070

[HPCwire](#) >> [Off the Wire](#)

July 20, 2009

## NEC Cluster Ranks High on Green500

Page: 1 of 3

[1](#) | [2](#) | [3](#) All »

DUSSELDORF, Germany, July 20 -- NEC demonstrates its international leadership in high performance computing (HPC) with an outstanding position in the recently published Green500 list. Installed at the High Performance Computing Center Stuttgart (HLRS), the new NEC LX-2400 HPC Cluster delivers an excellent value of 273 MFLOPS per watt, which is ahead of all other clusters equipped with Intel or AMD processors. This could be realized by highly efficient power supplies and diskless compute nodes.

"Energy-efficient systems and related cost savings are gaining increasing importance. With the new LX-Series Cluster NEC realized an optimal combination of low power consumption and high computing performance," explains Dr. Andreas Findling, product manager at NEC Deutschland GmbH.

The excellent ranking of the new system in the Green500 clearly shows the superiority of our solutions in terms of energy consumption compared to other systems in the list based on commodity processors.

Off the Wire

Aug 7, 2009

▶ [DOE Awards](#)  
Centers

▶ [Japan's](#)

Aug 6, 2009

▶ [Darkstrata](#)  
Innovation

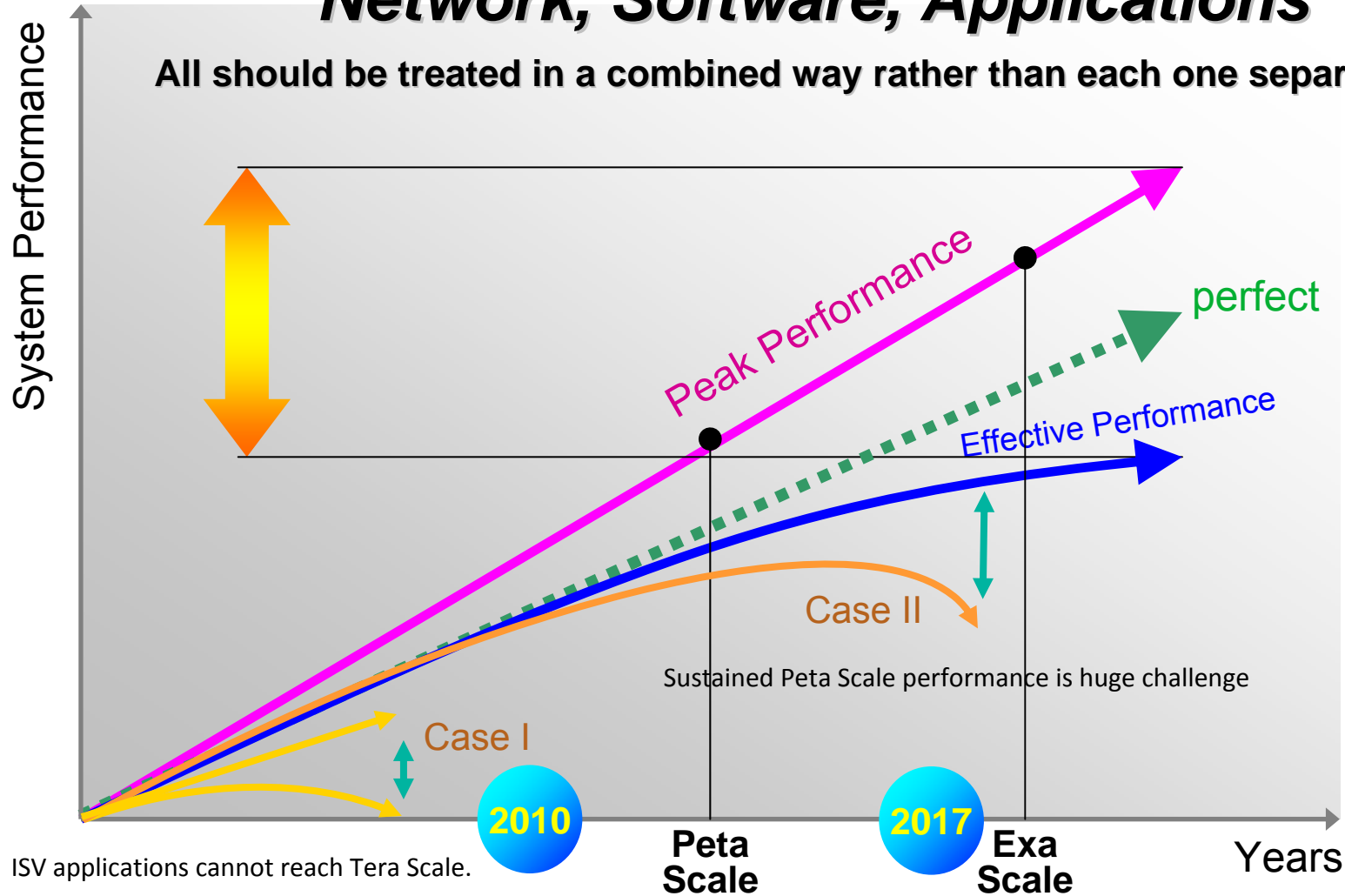
**The Green500 List**

**273 MFLOPS per Watt**

# Scalability Challenges for future systems

## *Network, Software, Applications*

All should be treated in a combined way rather than each one separately



# NEC's Supercomputing Approach

## NEC's Approach

- Get back to real HPC efficiency
  - „we expect 3% efficiency in 2011“
- Performance & Efficiency out of commodity components
- Achieve real HPC Efficiency – like with the classical way

## Taking x86 clusters to Capability Computing



**Computing Efficiency is THE key.**

# LX Series Network Features 2010

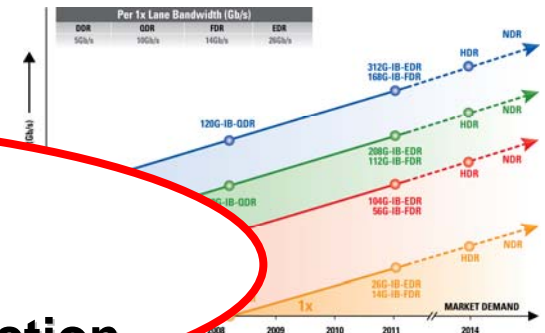
## HPC efficiency by optimized network performance

- Get an understanding of „real network performance“
  - Monitoring (**UFM**)
- Optimize network for application performance
  - Avoid static routing by topology awareness (**UFM & TARA**)
- Boost the performance of MPI collective operations
  - Making use of intelligent switches (**FCA – Fabirc Collectives Aclerator**)
- Optimize the InfiniBand software stack
  - Get away from usage of VERBS
- Intelligent way to share the network for I/O and MPI communication
  - Use QoS features for separation (**UFM**)

Empowered by Innovation

**NEC**

Total System Optimization  
by Optimizing Components  
and Optimal Component Interaction





Empowered by Innovation

Your HPC Solution provider

**NEC**



# About Voltaire (NASDAQ: VOLT)

- ▶ **Leading provider of Scale-out Data Center Fabrics**
  - Used by more than 30% of Fortune100 companies
  - Hundreds of installations of over 1000 servers
- ▶ **Addressing the challenges of HPC, virtualized data centers and clouds**
- ▶ **More than Half of Top500 InfiniBand Sites**
- ▶ **InfiniBand and 10GbE Scale-out Fabrics**

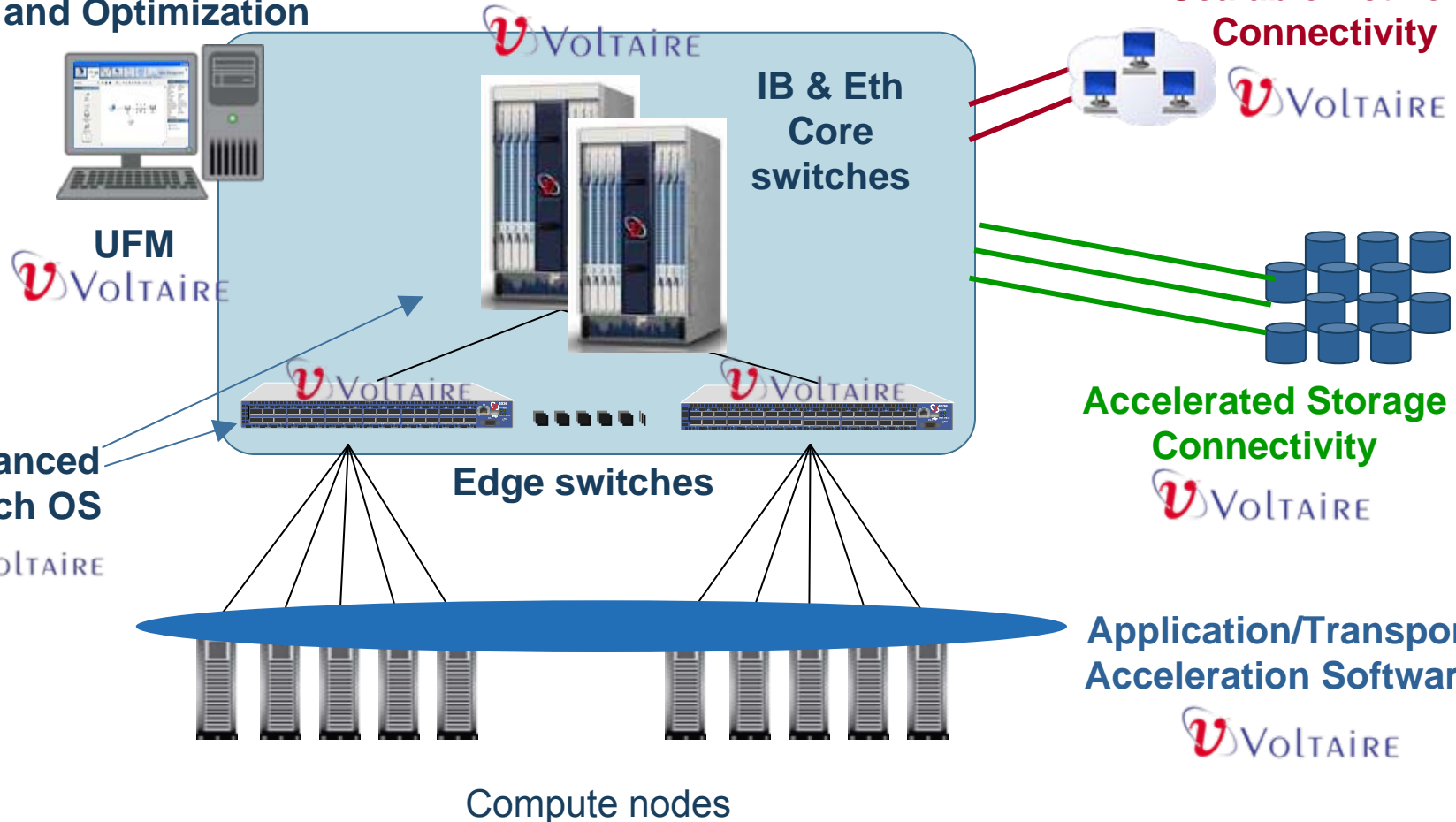
## End-to-End Scale-out Fabric Product Line



# Voltaire Products: End to End HPC Connectivity Solutions

**Advanced Fabric Management  
and Optimization**

**Scalable Network  
Connectivity**



**Accelerated Storage  
Connectivity**

**Application/Transport  
Acceleration Software**

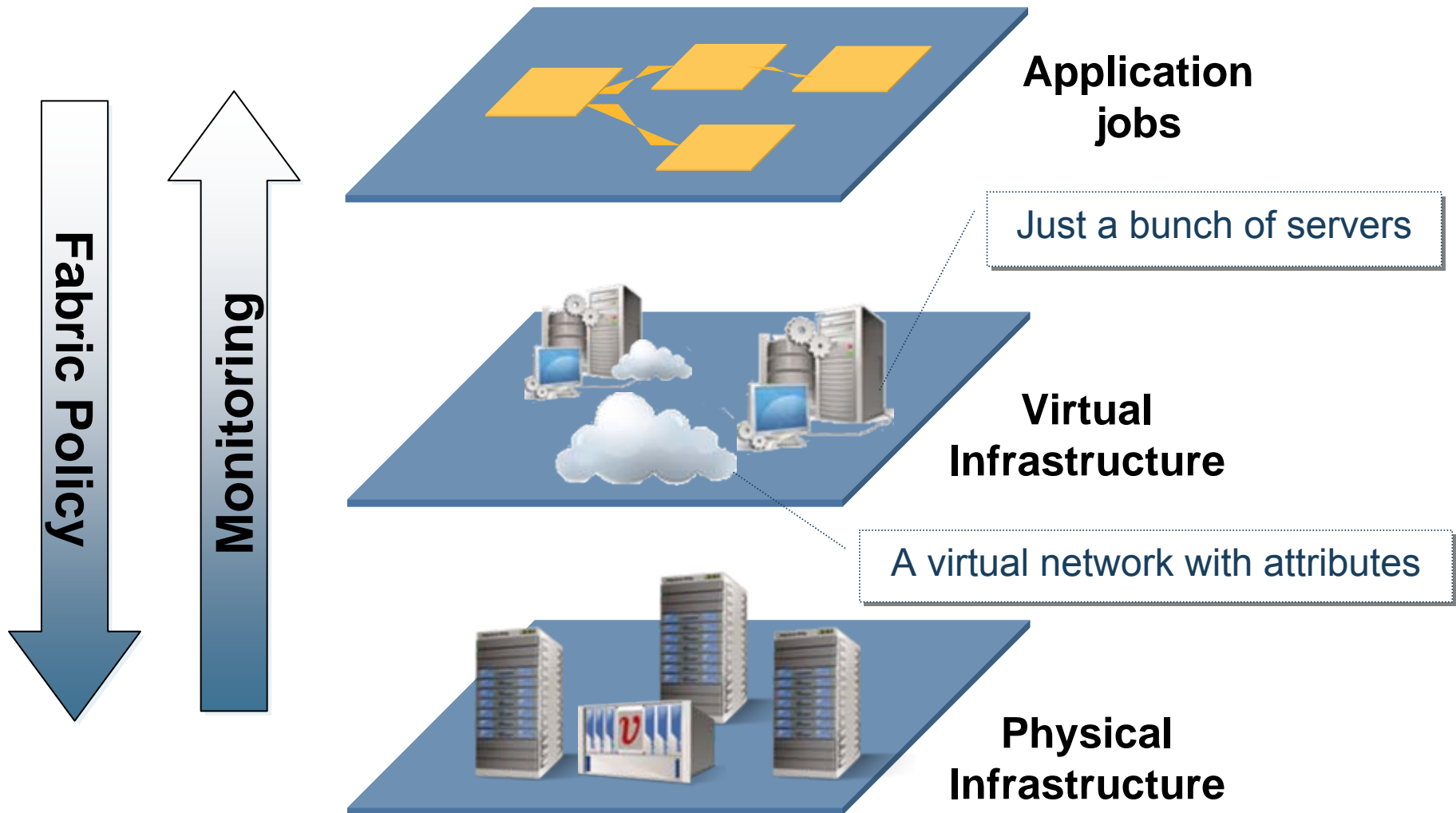
# I/O is the Bottleneck !

## *What shall we do ?*

**Add more switches, cables, servers, storage ?**

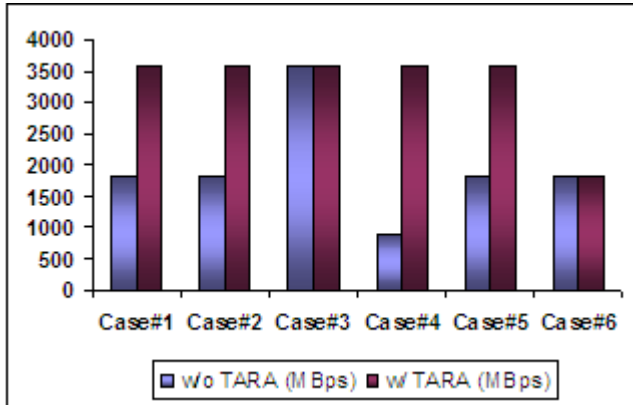
**Or perhaps eliminate the bottlenecks  
with software?**

# Application Correlated Fabric Management

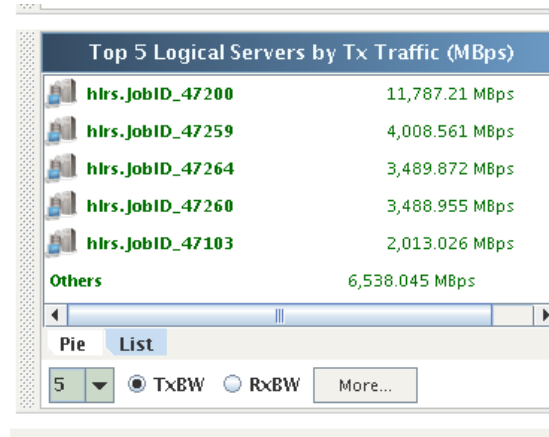


# Voltaire Unified Fabric Manager at Work

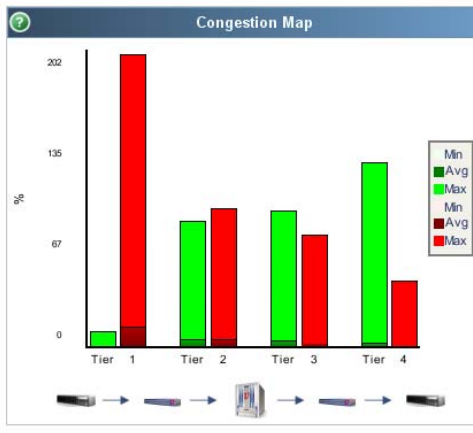
## Generating business value to customers



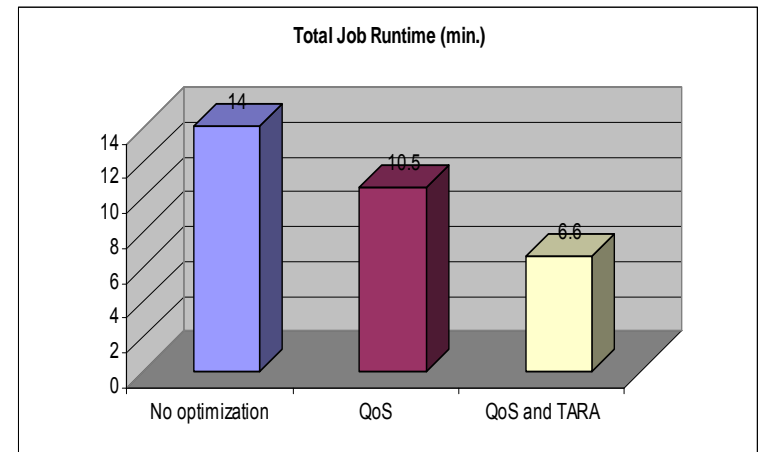
**UFM TARA improves customer performance by 200%**



**Admins can see which jobs consume their fabric resources**



**A global bank used UFM to detect sever congestions which impacted trading**

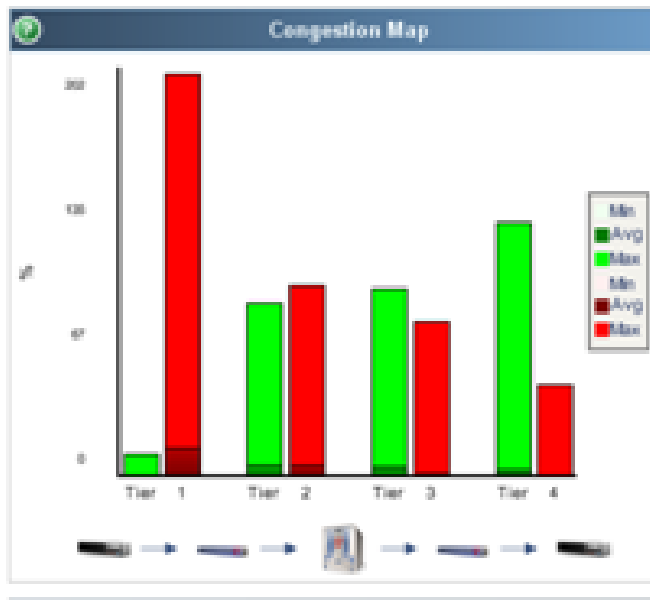


**UFM cuts customer job run-time, and provides differentiated services**

# Immediate Visibility

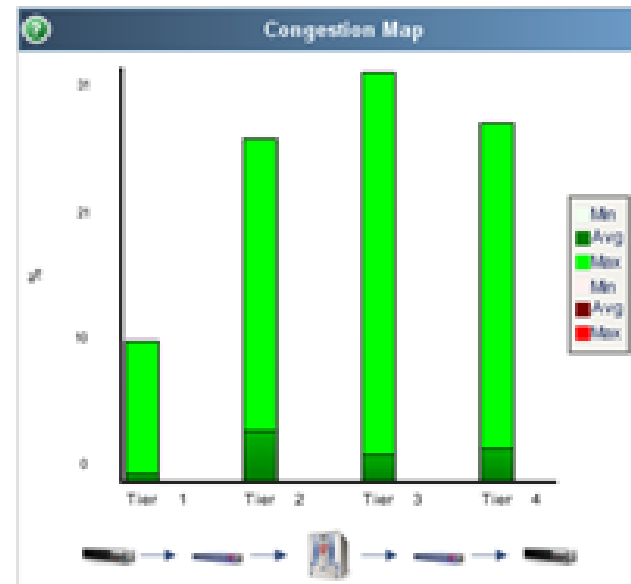
## Congestions

Before



Random Routing Cause Congestions

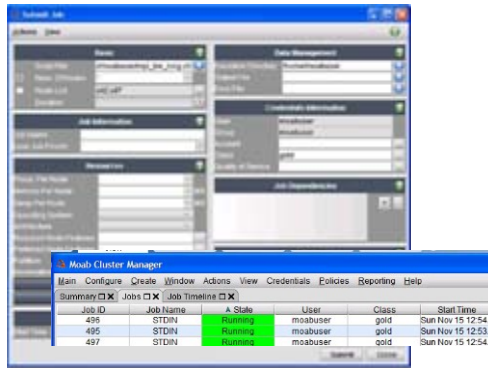
After



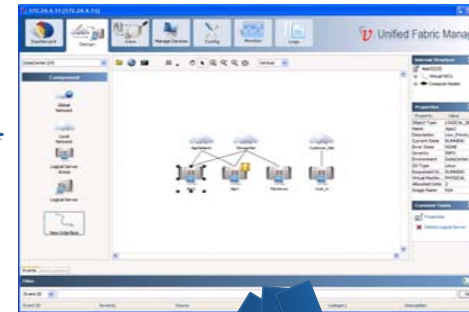
UFM Traffic Aware Routing™ Algorithm (TARA) Applied  
No Congestions

# UFM Integrated With Job Schedulers Dynamically Optimizing Jobs Fabric Utilization

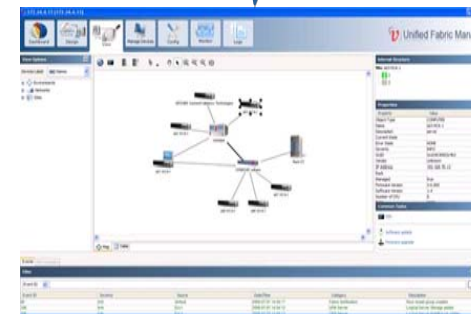
Job Submitted in Moab



Matching Jobs Automatically  
Created in UFM

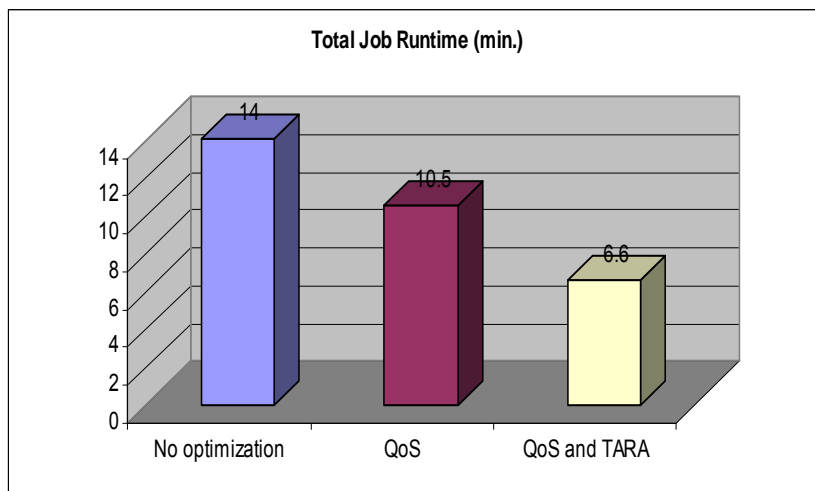
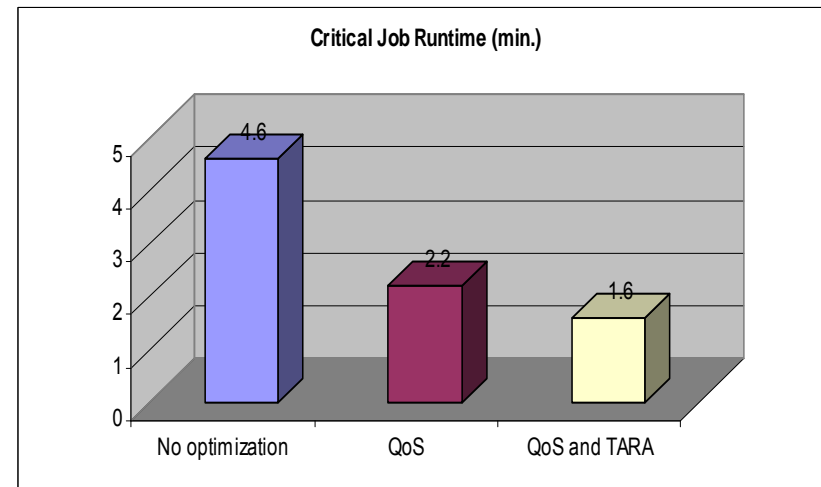
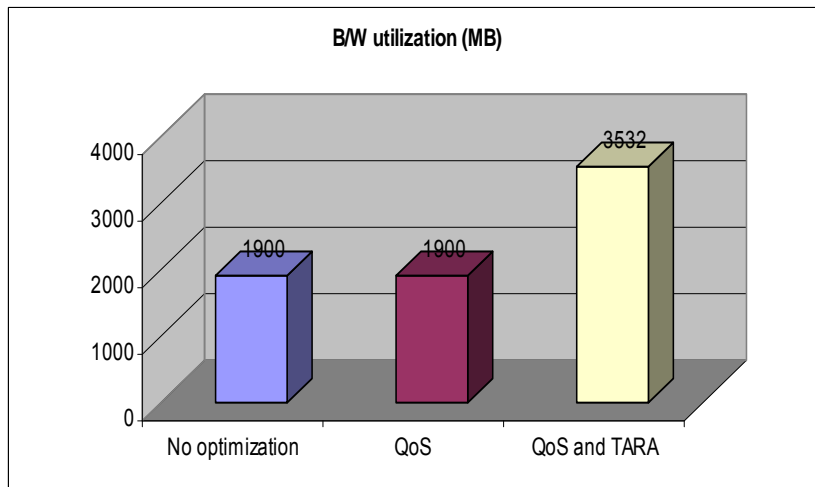


Application Level Monitoring  
& Optimization Measurements



Fabric-wide Policy Pushed to Match  
Application Requirements

# What Have We Achieved?



- ▶ **Overall job time reduced by 60%**
- ▶ ***Critical* job runs 3 times faster**
- ▶ **B/W utilization doubled**
- ▶ **Job level analysis**
- ▶ **Efficient troubleshooting**

## Recalculations of routes

- UFM offers TARA - the valuable feature of recalculation of the existing routes in the network for a new coming job
- Very important especially for blocking networks
- Integration with MOAB allows applying TARA, QoS and FCA dynamically
- MOAB is doing the placement → UFM is calculating the routes
- Next Steps:
  - Intelligent job placement: UFM to determine job placement **and** routes based on fabric topology
  - Even better would be in taking user information to optimize the parallel communication graph

## And the future?

- We expect PGAS parallel programming paradigms to come up
  - Coarray Fortran
  - UPC (Unified Parallel C)
  - Chapel
  - X10
  - Fraunhofer FVM (GPI)
- PGAS needs hardware support for running many short messages
  - One sided data access to remote memories (get and put)
  - Weak consistency needed
  - Efficient barriers for large and small subsets
  - Long series of short messages to and from many destinations