Infiniband Kopplung Heidelberg–Mannheim
Tests aktueller Komponenten und Pläne für den Ausbau

S. Richling, S. Friedel (Universität Heidelberg)
S. Hau, H. Kredel (Universität Mannheim)

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

UNIVERSITÄT
MANNHEIM

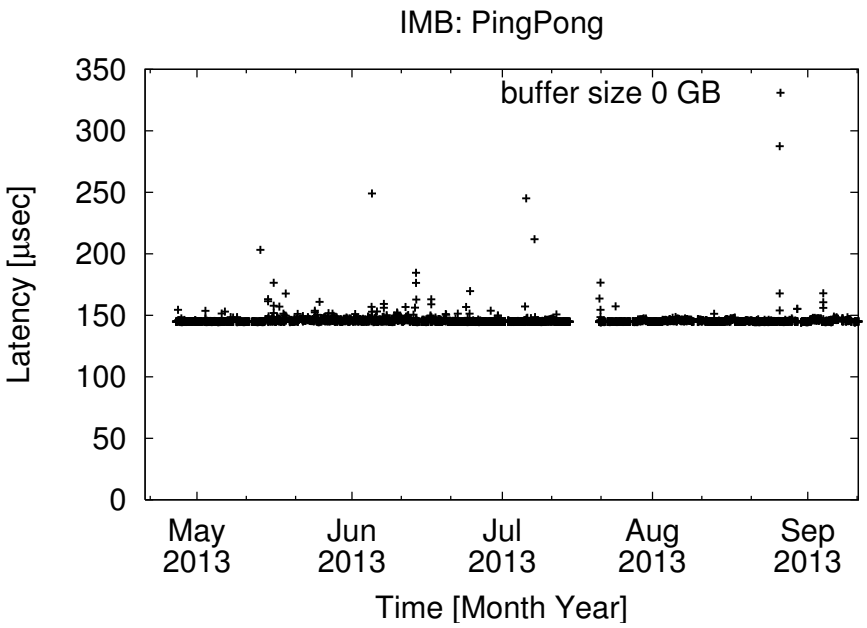
Förderträger:
Baden-Württemberg
MINISTERIUM FÜR WISSENSCHAFT, FORSCHUNG UND KUNST

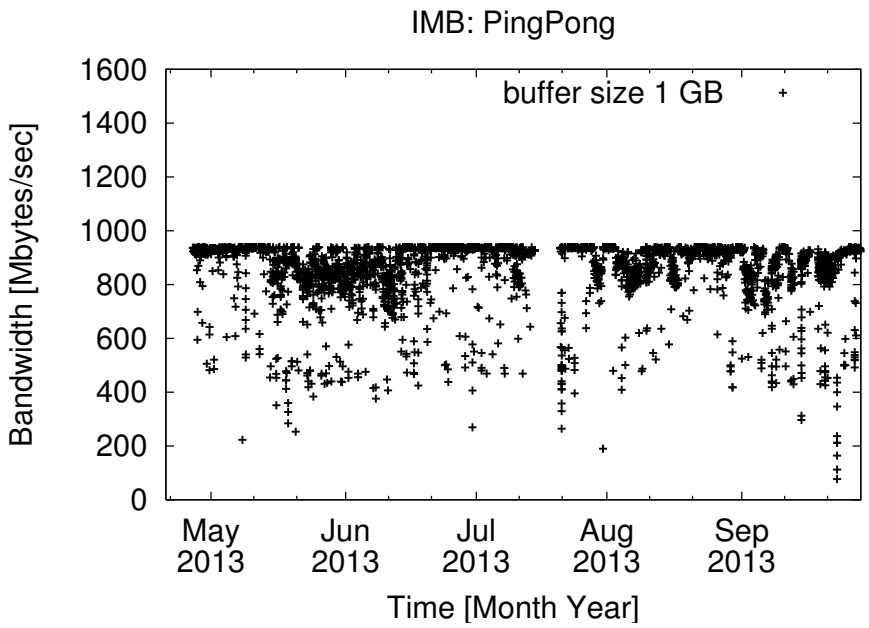
www.bwhpc-c5.de

Infiniband Kopplung Heidelberg–Mannheim
Derzeitiger Ausbau für bwGRiD

- Infiniband über Ethernet über Glasfaser
- Infiniband (20 Gbit/sec) $\leftrightarrow$ Ethernet (10 Gbit/sec) mit Obsidian Longbow
- stabil in Betrieb seit Juli 2009
  - eine Cluster-Administration mit einem Batchsystem
  - eine Benutzerverwaltung, die verschiedene Quellen integriert: LDAP (MA), AD (HD), Grid-Zertifikate, Shibboleth

Infiniband Kopplung Heidelberg–Mannheim
Latenz und Bandbreite bwGRiD



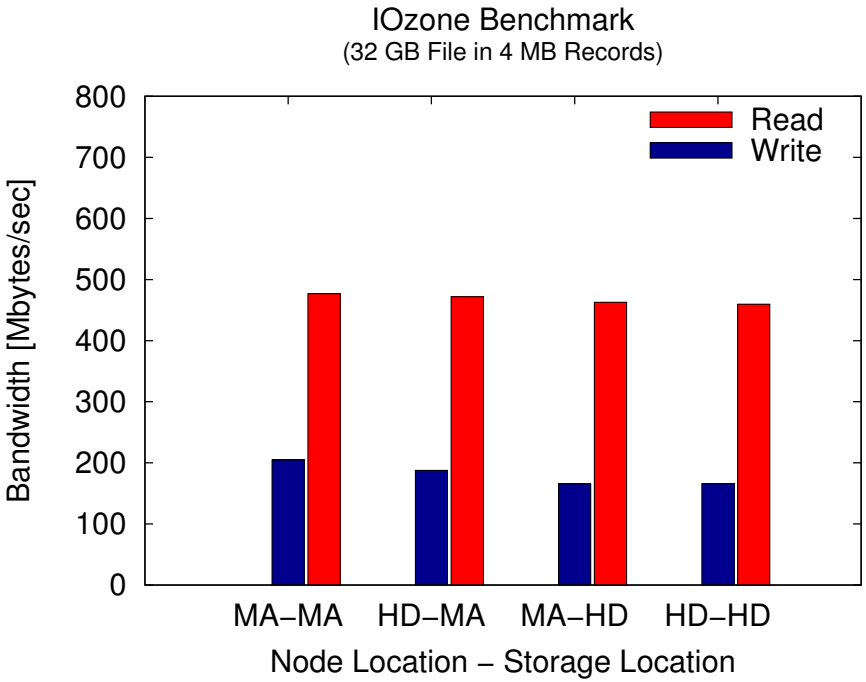


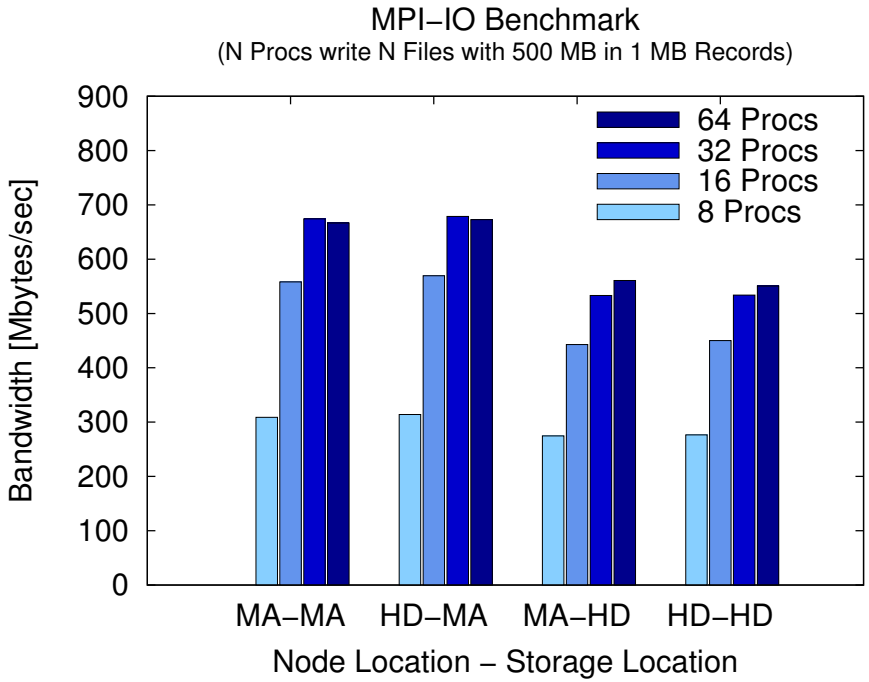
Lichtlaufzeit (28 km)	143 μsec
Latenz (Cluster)	+2 μsec
Latenz (MPI)	145 μsec

Bandbreite (MPI) $\leq$ 930 MByte/sec

- Latenz/Bandbreite nicht gut genug für standortübergreifende Jobs

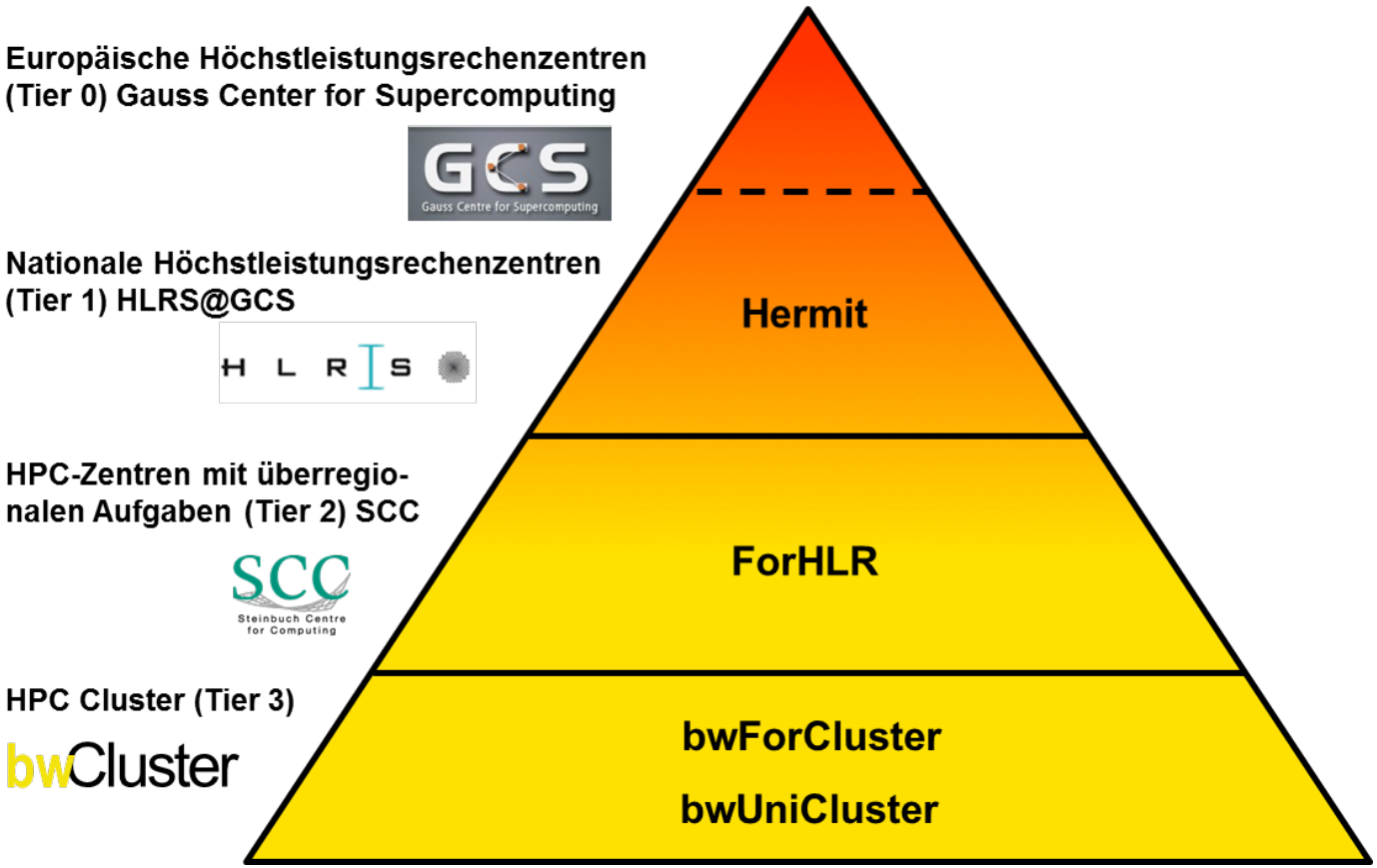
Infiniband Kopplung Heidelberg–Mannheim
IO Performance bwGRiD



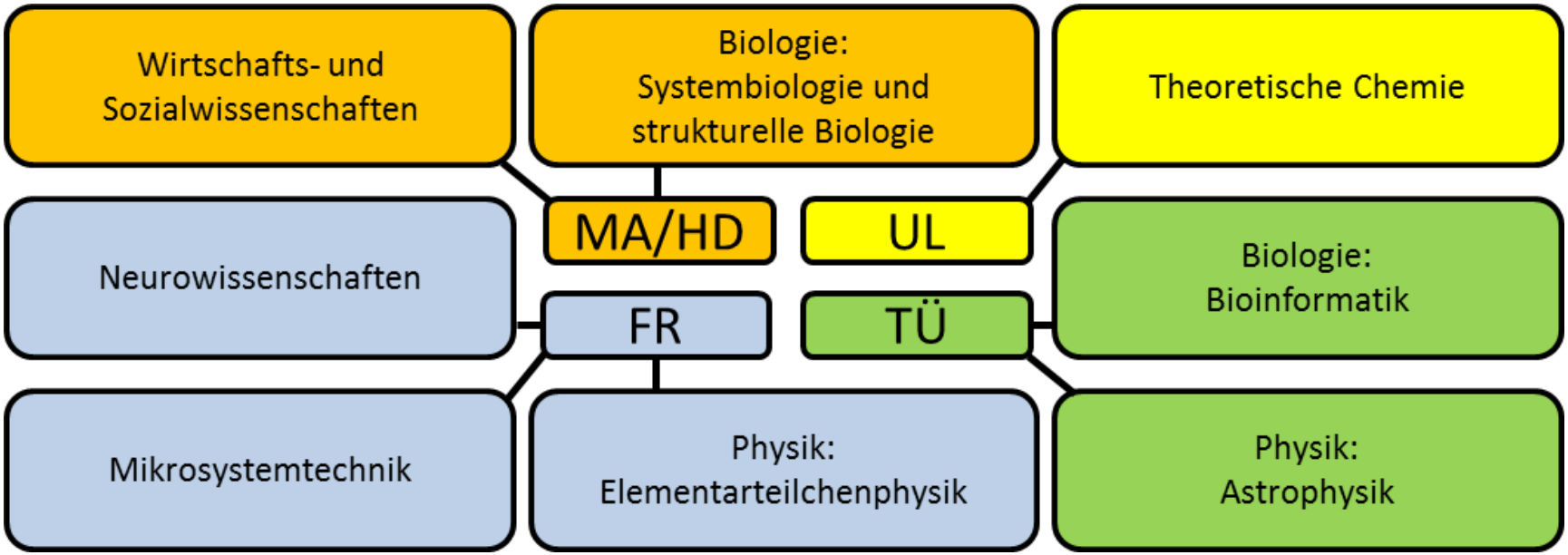


- Bandbreite ausreichend für Zugriff auf Storage-Systeme: Mannheim(MA) $HOME und Heidelberg(HD) $SCRATCH
- IO Performance nahezu unabhängig vom Zugriffsweg (Zugriff auf lokales oder entferntes Speichersystem)

bwHPC
Leistungspyramide



bwForCluster
Forschungsgebiete nach Standort



bwForCluster MLS&WISO
Standort Heidelberg/Mannheim

Forschungsschwerpunkte

- Molekulare Lebenswissenschaften
- Wirtschafts- und Sozialwissenschaften
- Wissenschaftliches Rechnen

Projektpartner

URZ UNIVERSITÄTS-RECHENZENTRUM

IWR Interdisciplinary Center for Scientific Computing

RECHENZENTRUM der rum UNIVERSITÄT MANNHEIM

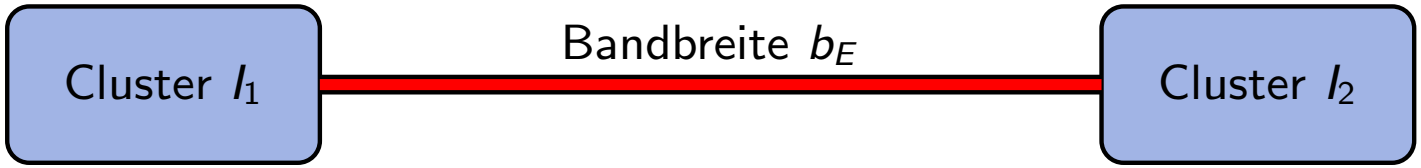
Inifiniband Kopplung Heidelberg–Mannheim
Überlegungen zum Ausbau für den bwForCluster

- Gründe für die Fortsetzung der Kopplung:
  - Bündelung und Ausbau von Kompetenzen (RUM, URZ, IWR)
  - Optimale Nutzung der Rechnerräume (Platz, Betriebskosten)
  - Erhöhte Verfügbarkeit und Ausfallsicherheit
- Netzwerk-Voraussetzungen für den Ausbau:
  - Kapazität der Glasfaser ausreichend (Dark Fiber)
  - Netzwerkkomponenten für Nutzung paralleler Kanäle nötig
  - 40 Gbit/sec Bandbreite möglich mit aktuell verfügbaren Netzwerkkomponenten Infiniband $\leftrightarrow$ Ethernet
- Durchsatzraten von aktuellen Speichersystemen sollen bedient werden (mehrere GByte/sec)
- Kosten für die Kopplung sollen im Rahmen bleiben
- Plan: Ausbau auf 4 x 40 Gbit/sec = 160 Gbit/sec

Performance Modell für verteilte Cluster

Kredel et al. 2012 (DOI 10.1007/s00450-012-0213-5)

Wenige Hardware-Parameter



n	Anzahl Kerne	n	Anzahl Kerne
l_{th}	Performance Kern	l_{th}	Performance Kern
b_l	Bandbreite	b_l	Bandbreite

Wenige Parameter für die Anwendung

$\#op$	Zahl der Rechenoperationen
$\#b$	Anzahl Bytes (Datenmenge)
$\#x$	Anzahl der ausgetauschten Bytes

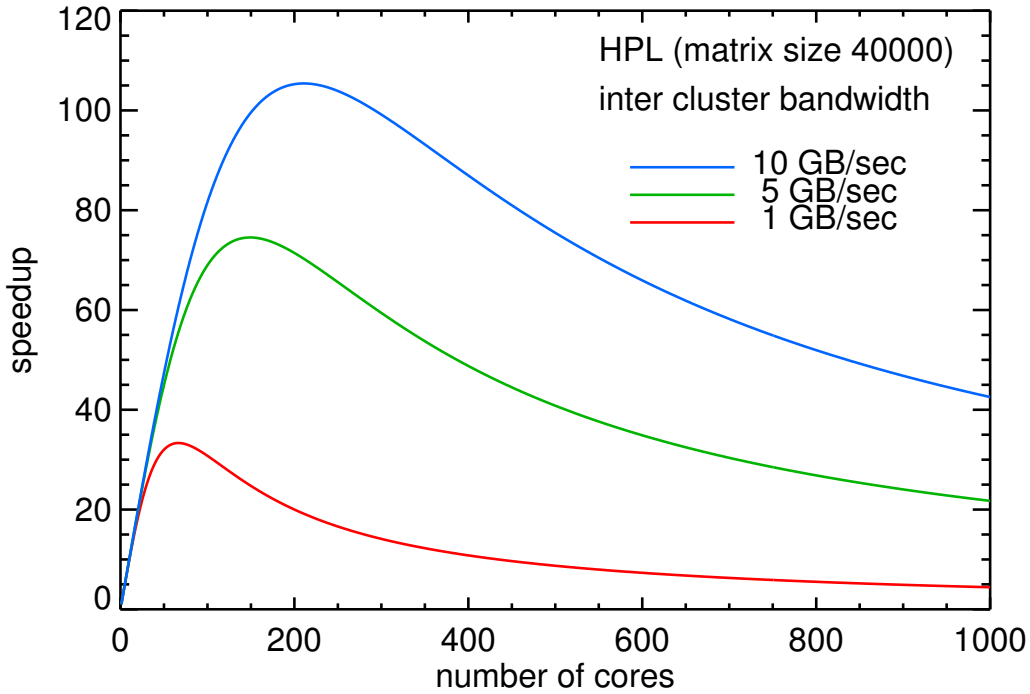
Skizze Performance Modell

Laufzeit auf einem Cluster: $t_1 =$ Rechenzeit + Kommunikation(b_l)

Laufzeit auf zwei Clustern: $t_2 = t_1$ (mit halber Load) + Kommunikation(b_E)

Ergebnis: Speedup für eine Anwendung für eine bestimmte Hardware

Standortübergreifende MPI Performance
Erwartungen für bwForCluster nach dem Performance Modell



- Ergebnis für kommunikationsintensive Anwendungen:
  - Lineare Skalierung bis etwa $n = 100$ für $b_E = 10$ GByte/sec.
  - 10-fache Bandbreite erhöht die Skalierbarkeit um Faktor $\sim$3.
- Standortübergreifende Jobs für bestimmte Anwendungen sinnvoll?

Infiniband über größere Entfernungen
Technische Möglichkeiten

- Mellanox MetroX Long Haul Series
  - Infiniband über Ethernet über Glasfaser
  - 56 Gbit/sec Infiniband $\leftrightarrow$ 4x10 Gbit/sec Ethernet
- Obsidian Longbow C-Series
  - Infiniband über Ethernet über Glasfaser
  - QDR Infiniband $\leftrightarrow$ 4x10 Gbit/sec Ethernet
- Wellenlängenmultiplexer zur Nutzung mehrerer Farbkanäle

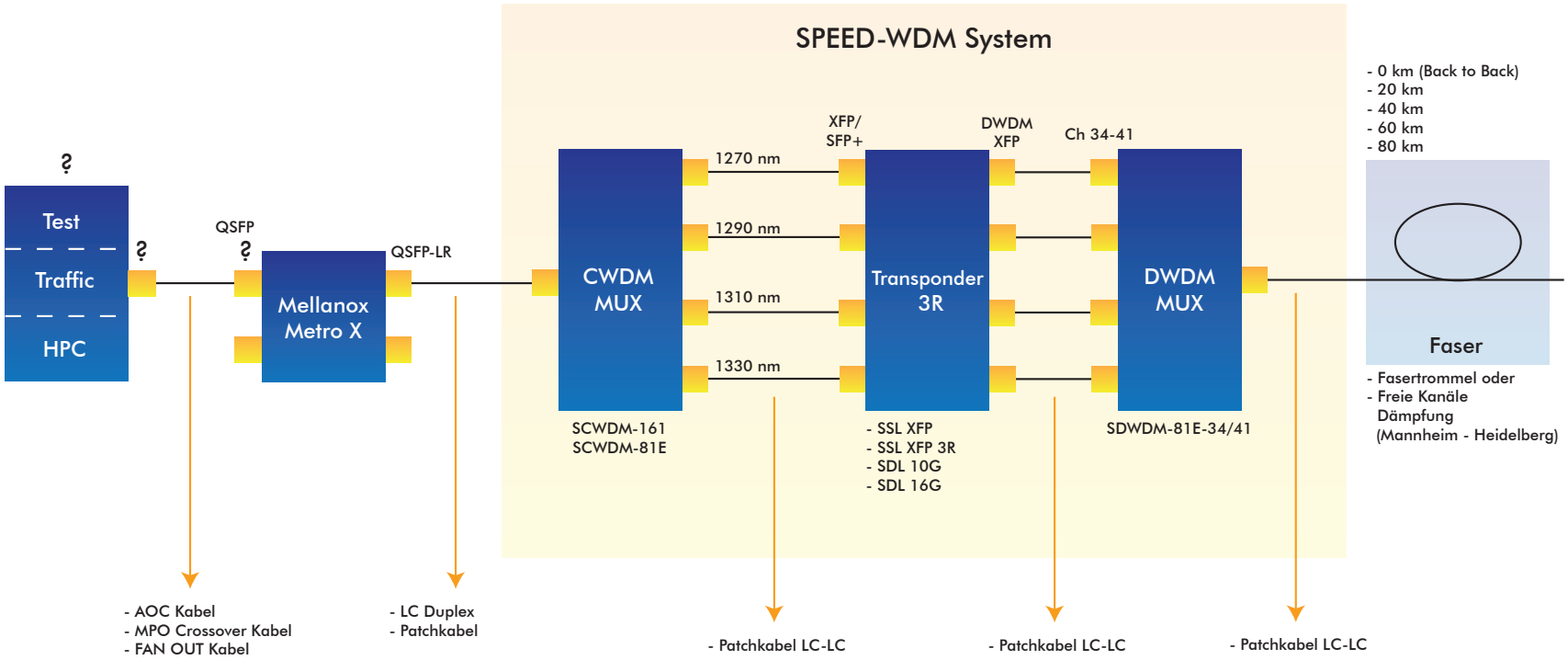
(Bildquelle: Pan DaCom)

Teststellung mit Mellanox
Aufbau (September 2013)

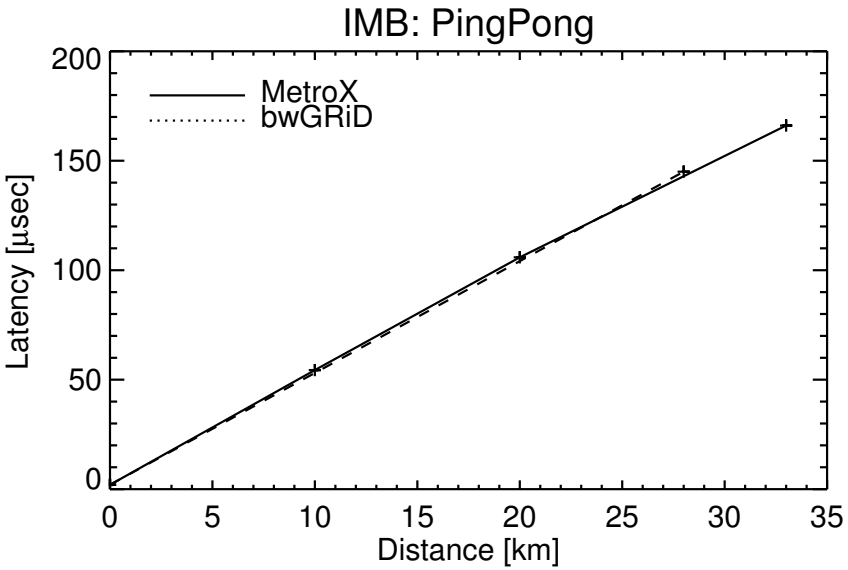
- Mellanox MetroX TX6100 Switche (bis 10 km)
- Pan Dacom DWDM System SPEED-OTS-5000
- HPC-Cluster Helics3a (IWR, Universität Heidelberg)
  - 32 Knoten mit 4 x 8 Core AMD Opteron
  - Mellanox 40G QDR single
- 4 Knoten verbunden über MetroX (1 x 40 Gbit/sec)
- Test-Entfernungen: 10 km, 20 km, 33 km
- Remote-Unterstützung durch Mellanox Entwickler

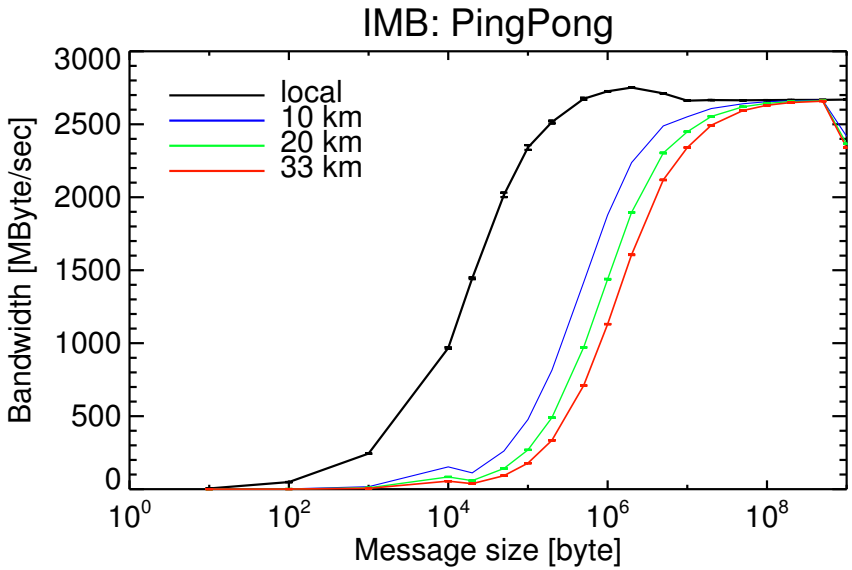
Teststellung mit Mellanox
Konfiguration (PoC)

40 Gbit/s Infiniband --> 4 x 10 Gbit/s --> DWDM



Teststellung mit Mellanox
Ergebnis Latenz und Bandbreite



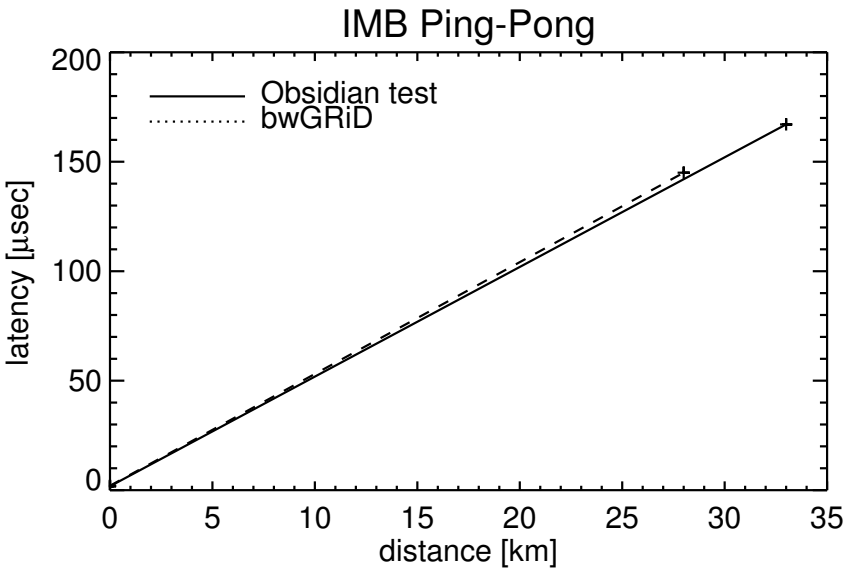


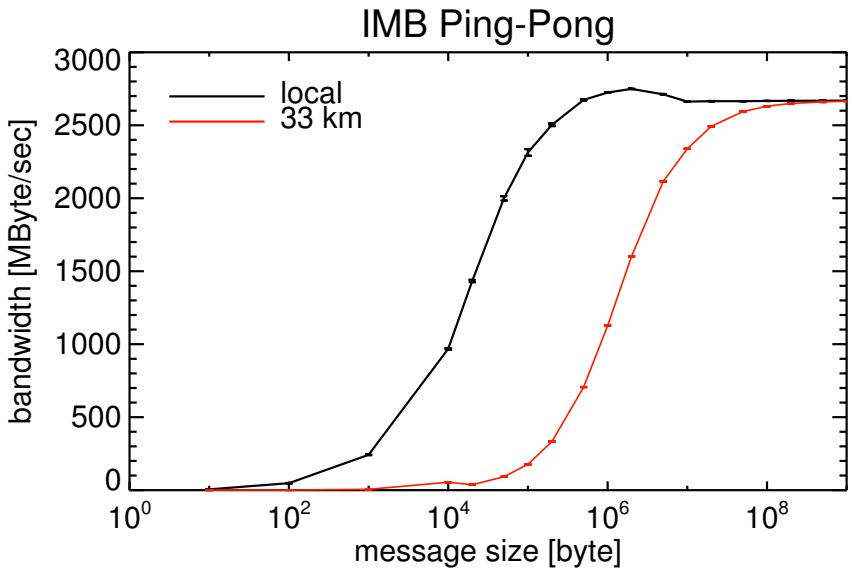
- Latenz für 33 km wie erwartet hoch.
- MPI-Bandbreite bei 1 x 40 Gbit/sec: 2.6 GByte/sec bis 33 km
- Erwartung MPI-Bandbreite bei 4 x 40 Gbit/sec: 10 GByte/sec

Teststellung mit Obsidian
Aufbau (April 2014)

- Obsidian Longbow C400
- Pan Dacom SPEED-CWDM 161
- HPC-Cluster Helics3a (IWR, Universität Heidelberg)
  - 32 Knoten mit 4 x 8 Core AMD Opteron
  - Mellanox 40G QDR single
- 4 Knoten verbunden über Obsidian Longbow (1 x 40 Gbit/sec)
- Test-Entfernung: 33 km

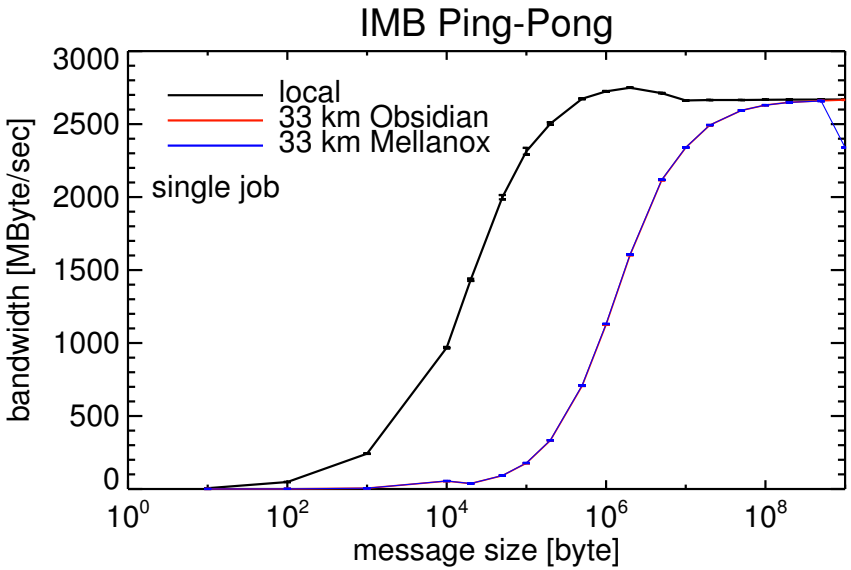
Teststellung mit Obsidian
Ergebnis Latenz und Bandbreite

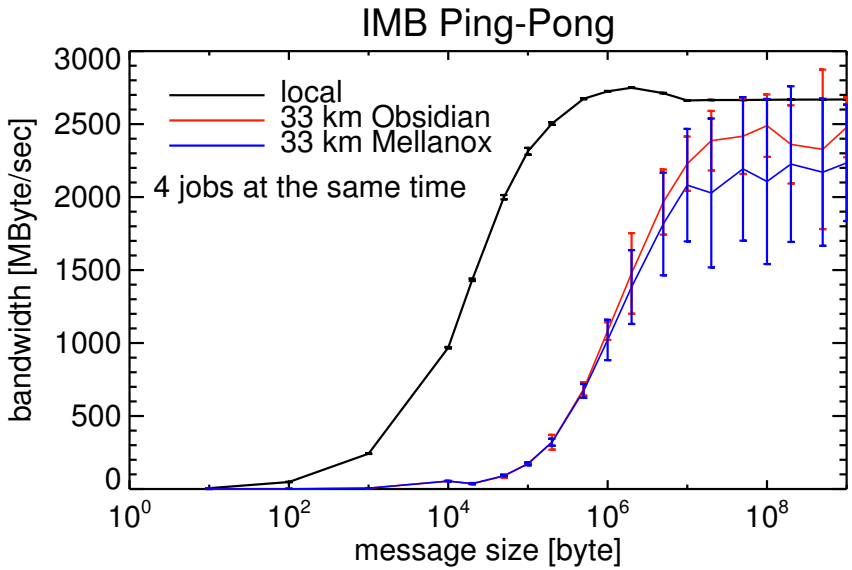




- Latenz für 33 km wie erwartet hoch.
- MPI-Bandbreite bei 1 x 40 Gbit/sec: 2.6 GByte/sec bis 33 km
- Erwartung MPI-Bandbreite bei 4 x 40 Gbit/sec: 10 GByte/sec

Vergleich der Teststellungen
Ergebnisse Bandbreite





- Software gleich: IMB 3.2, Intel Compiler 13.1.2, OpenMPI 1.6.4
- 1 Job: Bandbreite gleich, kleine Abweichung bei hoher Paketgröße
- 4 Jobs gleichzeitig: Bandbreite innerhalb der Fehlergrenzen gleich, Sättigung der Leitung erreicht

Zusammenfassung

- Infiniband Kopplung zum Betrieb eines verteilten Clusters
  - stabil und dauerhaft möglich
  - Fortführung des Konzeptes für neuen Cluster
- Teststellungen mit aktueller Technik
  - Funktionstests für 1 x 40 Gbit/sec mit Technik von Mellanox und Obsidian durchgeführt
  - Ergebnis: Beide Techniken sind für uns einsetzbar.
- Offene Punkte
  - Lasttest über längeren Zeitraum
  - Lastverteilung bei 4 x 40 Gbit/sec
  - Verhalten im Produktionsbetrieb