

bwGRID Heidelberg/Mannheim: Cluster-Betrieb mit einer Infiniband-Kopplung über 28 km

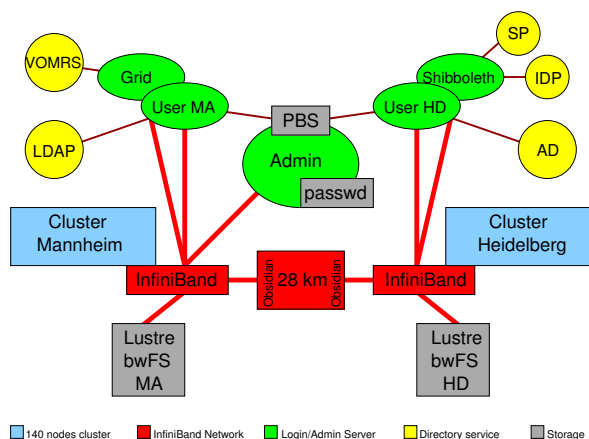
R. Bogus¹, S. Hau², H. Kredel², S. Richling¹

¹ Universität Heidelberg, Universitätsrechenzentrum, Im Neuenheimer Feld 293, 69120 Heidelberg

² Universität Mannheim, Rechenzentrum, L15,1-6, 68131 Mannheim

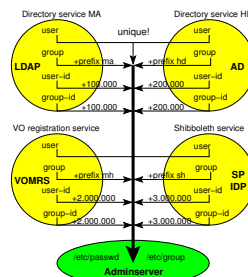
bwGRID Cluster Heidelberg/Mannheim

Konfiguration



Betrieb

Provisionierung



Authentifizierung

- Zugang mit Account von Mannheim (LDAP)
- Zugang mit Account von Heidelberg (AD)
- Zugang mit Grid-Zertifikat (VOMRS)
- Zugang über bwIDM (Shibboleth SP/IDP)

Scheduling

- Jobs laufen entweder in MA oder in HD
- Keine Jobs über die Standorte hinweg

Vorteile

- Gemeinsame Administration aller Knoten
- Höhere Auslastung des Gesamtsystems

Referenzen

S. Richling, S. Hau, H. Kredel, H.-G. Kruse, Operating Two InfiniBand Grid Clusters over 28 km Distance, International Journal of Grid and Utility Computing 2011, Vol. 2, No.4, pp. 303 - 312

S. Richling, S. Hau, H. Kredel, H.-G. Kruse, A Long-distance InfiniBand Interconnection between two Clusters in Production use, Proceedings Supercomputing 2011, November 12-18, 2011, Seattle, Washington, USA

Kopplung

Beschaffung und Aufbau

- Antrag in 2008
- Beschaffung und Aufbau bis Mai 2009
- In Betrieb seit Juli 2009
- Technik: InfiniBand über Glasfaser (Obsidian Longbow)



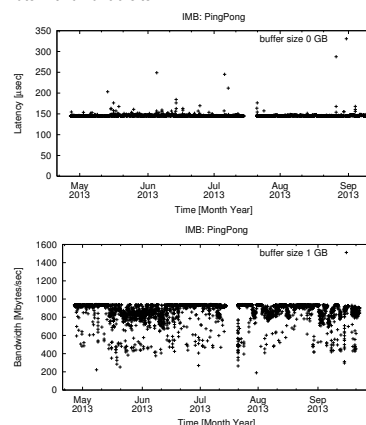
Zur Latenz und Bandbreite

- Entfernung MA-HD ist 28 km → Lichtlaufzeit ist 143 μ sec
- Latenz: 145 μ sec = Lichtlaufzeit + 2 μ sec lokale Latenz
- MPI-Bandbreite: etwa 930 MByte/sec

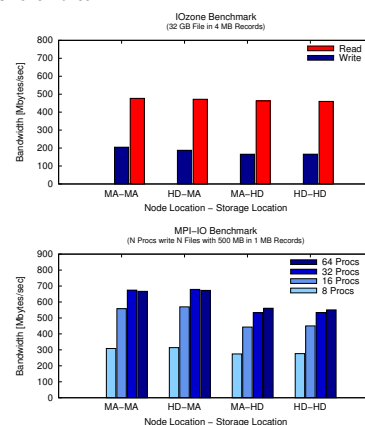
Erfahrungen mit der Kopplung

- Kopplung ist sehr stabil
- Bandbreite ausreichend für Zugriff auf Storage-Systeme
- IO Performance nahezu unabhängig vom Zugriffsweg

Latenz und Bandbreite



IO Performance



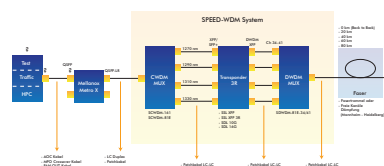
bwForCluster Heidelberg/Mannheim

Geplante Technik: Mellanox MetroX

- InfiniBand über MetroX
- Geplanter Ausbau: 4 x 40 Gbit/sec = 160 Gbit/sec

Konfiguration (PoC, September 2013)

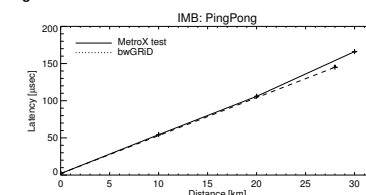
40 Gbit/s InfiniBand → 4 x 10 Gbit/s → DWDM



Test-Setup

- HPC-Cluster: Helios3a (IWR, Universität Heidelberg)
- Netzwerk: Mellanox QDR
- Einige Knoten verbunden über MetroX (1 x 40 Gbit/sec)
- Test-Entfernungen: 10 km, 20 km, 30 km

Ergebnis: Latenz



Ergebnis: Bandbreite

- MPI-Bandbreite bei 1 x 40 Gbit/sec: 2.6 GByte/sec bis 30 km
- Erwartung MPI-Bandbreite bei 4 x 40 Gbit/sec: 10 GByte/sec

