Performance Analysis and Prediction for distributed homogeneous Clusters

Heinz Kredel, Hans-Günther Kruse, Sabine Richling, Erich Strohmaier

IT-Center, University of Mannheim, Germany
IT-Center, University of Heidelberg, Germany
Future Technology Group, LBNL, Berkeley, USA

ISC'12, Hamburg, 18. June 2012

Outline

1. Background and Motivation
   - D-Grid and bwGRiD
   - bwGRiD Mannheim/Heidelberg
   - Next generation bwGRiD
2. Performance Modeling
   - The Roofline Model
   - Analysis of a single Region
   - Analysis of two identical interconnected Regions
   - Application to bwGRiD
3. Conclusions

D-Grid and bwGRiD

- bwGRiD Virtual Organization (VO)
  - Community project of the German Grid Initiative D-Grid
  - Project partners are the Universities in Baden-Württemberg
- bwGRiD Resources
  - Compute clusters at 8 locations
  - Central storage unit in Karlsruhe
- bwGRiD Objectives
  - Verifying the functionality and the benefit of Grid concepts for the HPC community in Baden-Württemberg
  - Managing organizational, security, and license issues
  - Development of new cluster and Grid applications

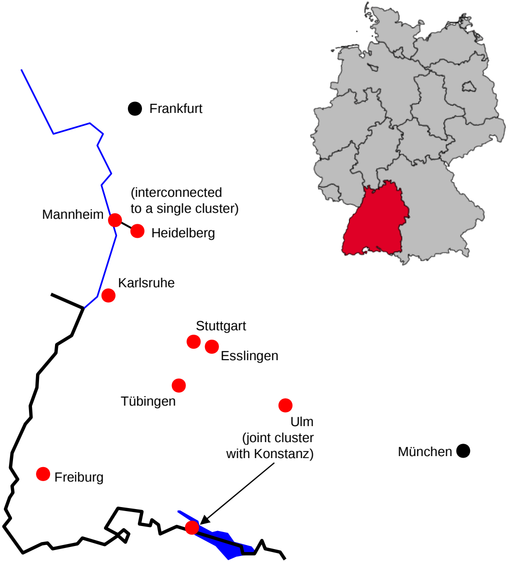
bwGRiD – Resources

Compute Cluster

Site	Nodes
Mannheim	140
Heidelberg	140
Karlsruhe	140
Stuttgart	420
Tübingen	140
Ulm/Konstanz	280
Freiburg	140
Esslingen	180
Total	1580

Central Storage

with backup	128 TB
without backup	256 TB
Total	384 TB



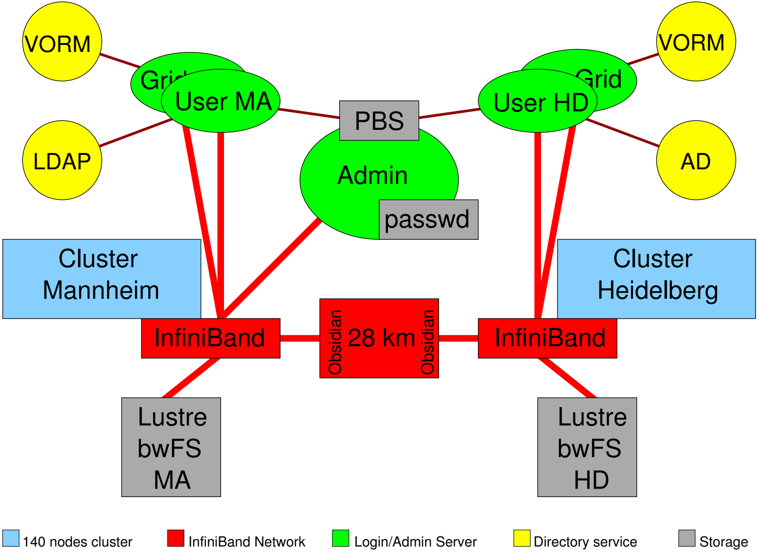
bwGRiD MA/HD – Hardware

Hardware	Mannheim	Heidelberg	total
Blade Center	10	10	20
Blades (Nodes)	140	140	280
CPUs (Cores)	1120	1120	2240
Login Server	2	2	4
Admin Server	1	–	1
Infiniband Switches	1	1	2
HP Storage System	32 TB	32 TB	64 TB

Blade Configuration

- 2 Intel Xeon CPUs, 2.8 GHz (each CPU with 4 Cores)
- 16 GB Memory
- 140 GB hard drive (since January 2009)
- Gigabit-Ethernet (1 Gbit)
- Infiniband Netzwork (20 Gbit)

bwGRiD MA/HD – Overview



bwGRiD MA/HD – Interconnection

Network Technology

- InfiniBand over Ethernet over fibre optics (28 km)
- 2 Obsidian Longbow (150 TEUR)

MPI Performance

- Latency is high: 145 μsec $=$ 143 μsec light transit time $+$ 2 μsec
- Bandwidth is as expected: 930 MB/sec (local 1200-1400 MB/sec)

Operating Considerations

- Operating the two clusters as single system image
- Fast InfiniBand interconnection to the storage systems
- MPI performance not sufficient for all kinds of parallel jobs
  $\rightarrow$ Keep all nodes of a job on one side

Next generation bwGRiD

Questions

- What bandwidth is required to allow all parallel jobs running accross two cluster regions?
- Is the expected bandwidth for the new system sufficient?
- Is there an optimal size for a cluster region?

Performance Charateristics	bwGRiD 1	bwGRiD 2
Bandwidth between two nodes	1.5 GByte/sec	6 GByte/sec
Bandwidth between two regions	1.0 GByte/sec	15 – 45 GByte/sec
Performance of a single core	8.5 GFlop/sec	10 – 16 GFlop/sec

Outline

1. Background and Motivation
   - D-Grid and bwGRiD
   - bwGRiD Mannheim/Heidelberg
   - Next generation bwGRiD
2. Performance Modeling
   - The Roofline Model
   - Analysis of a single Region
   - Analysis of two identical interconnected Regions
   - Application to bwGRiD
3. Conclusions

EECS Electrical Engineering and Computer Sciences

Basic Roofline

BERKELEY PAR LAB

- Performance is upper bounded by both the peak flop rate, and the product of streaming bandwidth and the flop:byte ratio

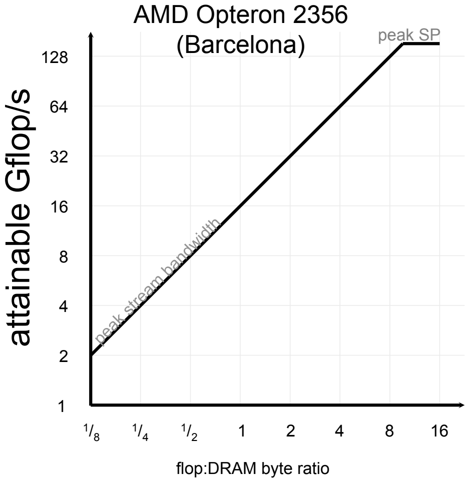
$$\text{Gflop/s} = \min \begin{cases} \text{Peak Gflop/s} \\ \text{Stream BW * actual flop:byte ratio} \end{cases}$$

1

EECS Electrical Engineering and Computer Sciences

Roofline model for Opteron
(adding ceilings)

BERKELEY PAR LAB



- Peak roofline performance
- based on manual for **single precision peak**
- and a hand tuned stream read for bandwidth

2

A Performance Model based on the Roofline Model

Roofline Principles:

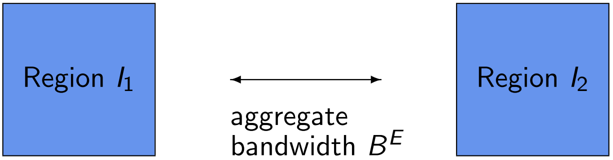
- Bottleneck Analysis
- Bound by Peak Flop and Measured Bandwidth

The following steps will be used to develop a performance model for single and multiple regions:

- Transform basics scales to dimensionless quantitates to arrive at universal scaling law
- Assume optimal floating-point operations and scaling with system size
- Introduce effective bandwidth scaling with system size
- Formulate result with dimensionless code-to-system balance factors

Performance Model – Overall System Abstraction

Hardware



number of cores n	number of cores n
core performance l_{th}	core performance l_{th}
bandwidth b^l	bandwidth b^l

Application (Load)

$\#op$ number of arithmetic operations performed on

$\#b$ number of bytes (data)

Analysis of a single Region

Total time =

Computation time + Communication time

Total time with ideal floating-point operations:

$$t_V \sim \begin{cases} \frac{\#op}{d^s} + \frac{\#b}{b^l} \\ \max\left(\frac{\#op}{d^s}, \frac{\#b}{b^l}\right) \end{cases} \geq \begin{cases} \frac{\#op}{d^{\mathrm{th}}}\left(1 + \frac{\#b}{b^l}\frac{d^{\mathrm{th}}}{\#op}\right) & \text{additive} \\ \frac{\#op}{d^{\mathrm{th}}} \max\left(1, \frac{\#b}{b^l}\frac{d^{\mathrm{th}}}{\#op}\right) & \text{overlapping} \end{cases}$$

Identify a **code-to-system balance factor** x based on:

a: Arithmetic intensity (roofline model, Williams et al. 2009)

a^*: Operational balance ('architectural intensity'):

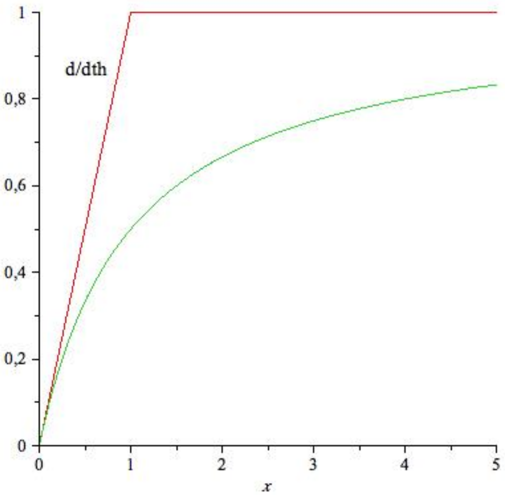
$$x = \frac{a}{a^*} = \frac{\#op}{\#b}\frac{b^l}{d^{\mathrm{th}}} = \frac{\#op}{d^{\mathrm{th}}}\frac{b^l}{\#b}$$

Throughput:

$$d = \frac{\#op}{t_V} \leq \begin{cases} d^{\mathrm{th}} \frac{x}{x+1} & \text{additive} \\ d^{\mathrm{th}} \min(1, x) & \text{overlapping} \end{cases}$$

Single Region – Throughput

Throughput d for additive (green) and overlapping (red) concepts.



Single Region – Speed-up

Ideal floating-point $d^{\mathrm{th}} = n \cdot l^{\mathrm{l}}$ and
Effective bandwidth scaling $z = \frac{b^l}{b^l_0}$ with a reference bandwidth b^l_0 gives:

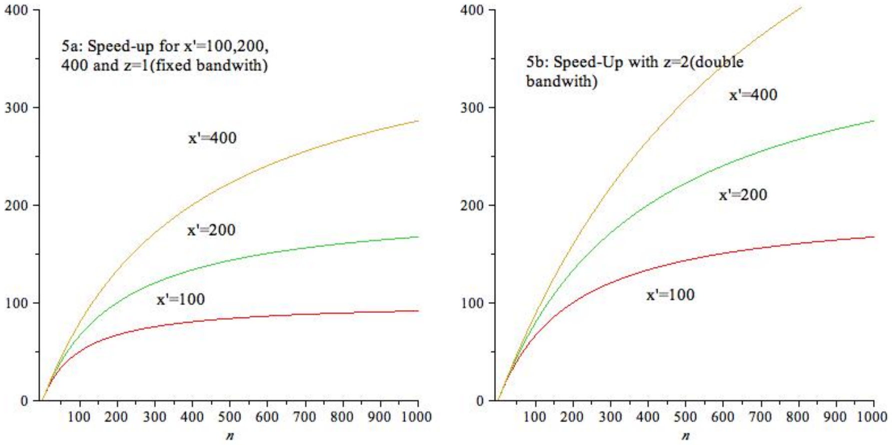
$$x = \frac{\#op}{\#b} \cdot \frac{b^l}{d^{\mathrm{th}}} = \frac{1}{n} \cdot \frac{\#op}{\#b} \cdot \frac{b^l_0}{l_{\mathrm{th}}} \cdot \frac{b^l}{b^l_0} = \frac{x' \cdot z}{n}$$

where x' is the balance factor of the core (or node, unit, ...)
Parallel Speed-up is then:

$$Sp = \frac{d(n)}{d(1)} = \frac{1 + x'z}{1 + \frac{x'z}{n}} \to 1 + x'z \Big|_{n \to \infty}$$

Single Region – Speed-up

Speed-up Sp for different values x' and z.



Analysis of two interconnected Regions

Total time =

Time (1 region, 1/2 comp. load) + Communication time between regions

Total time for $\#x$ bytes and channel bandwidth B^E :

$$t_V \sim t_V^{(1)} + \#x/B^E$$

$$t_V \geq \frac{(\#op/2)}{d^{\mathrm{th}}} \left(1 + \frac{a^*}{a}\right) + \frac{\#x}{B^E}$$

Throughput:

$$d \leq 2d^{\mathrm{th}} \frac{1}{1 + \frac{a^*}{a} + 2\frac{d^{\mathrm{th}}}{B^E}\frac{\#x}{\#op}}$$

Two Regions – Speed-up

Balance factors within (x') and between regions (y'):

$$x = \frac{a}{a^*} = \frac{x'}{n} \qquad y = \frac{\#op}{\#x}\frac{B^E}{2d^{\mathrm{th}}} = \frac{1}{2}\frac{x'}{n}\left(\frac{\#b}{\#x}\right)\left(\frac{B^E}{b^I}\right) = \frac{y'}{n}$$

Interconnection is a shared medium with a constant aggregate bandwidth B^E and an effective load factor $p(n)$:

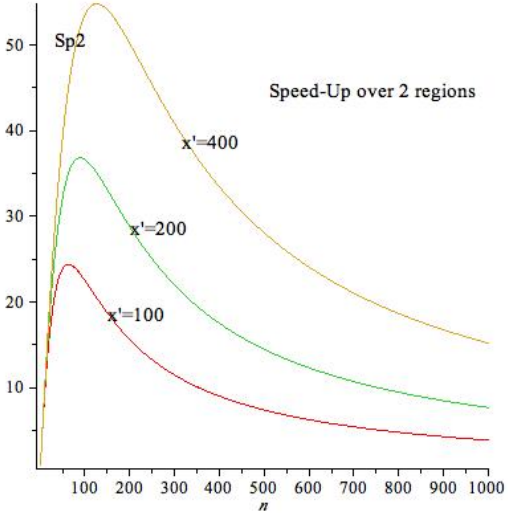
$$b^E = \frac{B^E}{p(n)}$$

This gives for the overall Speed-up:

$$Sp2 = \frac{x' + y' + x'y'}{p(n)x' + y' + \frac{x'y'}{n}} \to 0\Big|_{n\to\infty}$$

Two Regions – Speed-up

Speed-up $Sp2$ for different values of x'.



Two Regions – Speed-up

Focus on application and interconnection bandwidth:

$$z' = \frac{2y'}{x'} = r \cdot z'' \quad \text{with} \quad r = \frac{\#b}{\#x} \quad \text{and} \quad z'' = \frac{b^E}{b^I}$$

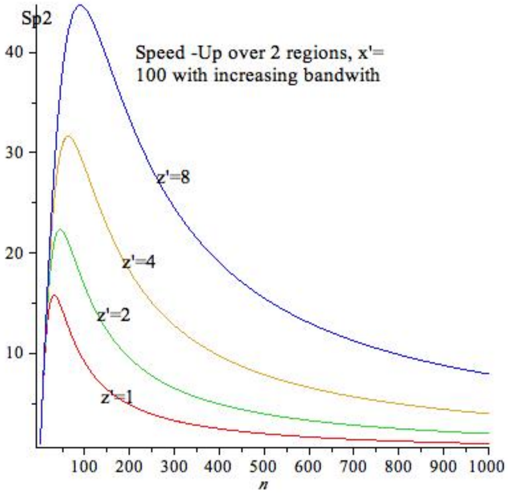
z' is the ratio between balance factors 'between regions' to 'between cores' and should be as large as possible

Overall Speed-up can be rewritten as:

$$Sp2 = \frac{2 + (1 + x')z'}{2p(n) + (1 + \frac{x'}{n})z'} \leq \frac{x'z'}{2p(n) + \frac{x'z'}{n}}$$

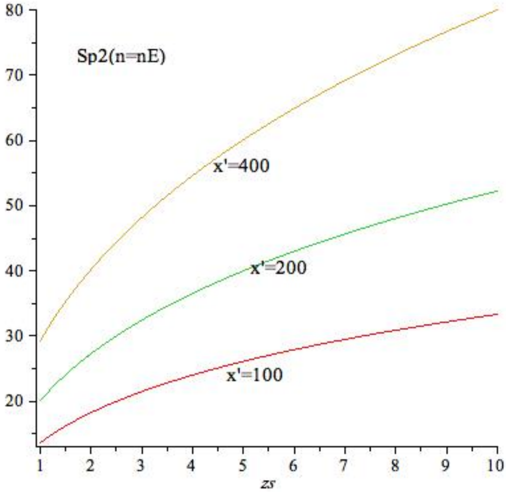
Two Regions – Speed-up

Speed-up $Sp2$ for $x' = 100$ with increasing bandwidth b^E (and consequently z') and an assumed $p(n) = \frac{n}{20}$.



Two Regions – Max. Speedup

Value of the maximum speed-up of $Sp2$ for linear $p(n) = \alpha n$ over bandwidth z'.



Application to bwGRiD

Performance Charateristics	bwGRiD 1	bwGRiD 2
Bandwidth between two nodes b^I	1.5 GByte/sec	6 GByte/sec
Bandwidth between two regions B^E	1.0 GByte/sec	15 GByte/sec
Performance of a single core l_{th}	8.5 GFlop/sec	10 GFlop/sec

Reference Bandwidth: $b_0^I = 1.0$ GByte/sec

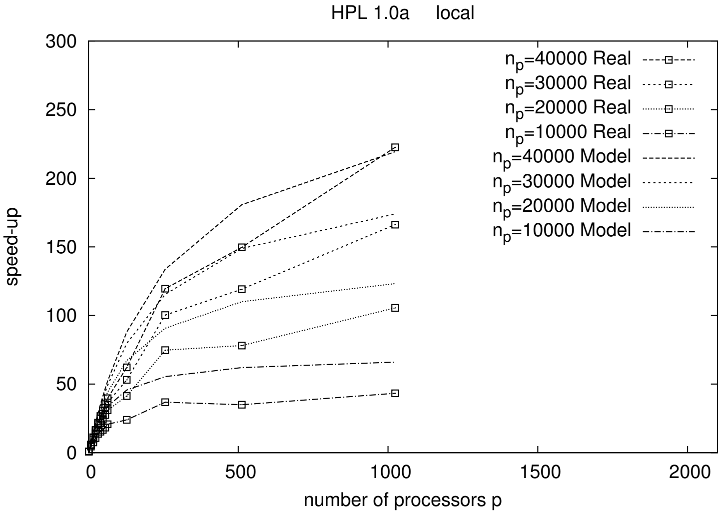
Application = LinPack:

$$n_p = 10000, 20000, 30000, 40000$$

$$\#op \sim \frac{2}{3} n_p^3 \qquad \text{and} \qquad \#b \sim 2 n_p^2 \cdot w$$

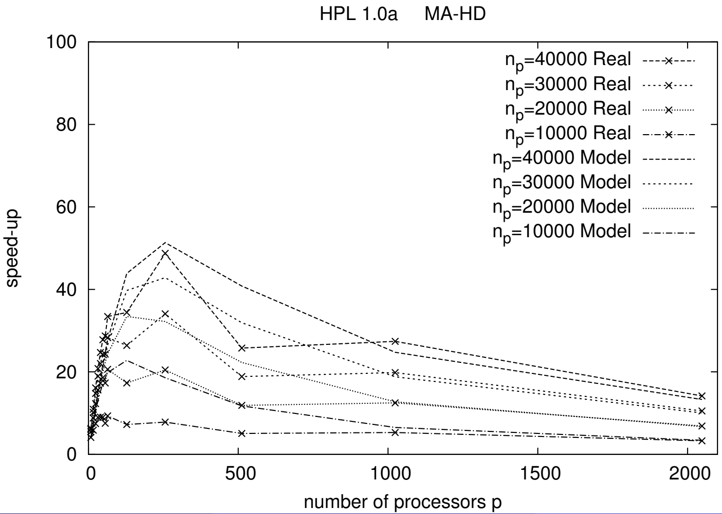
bwGRiD – Single Region

Speed-up comparison of measurements and model for one region.



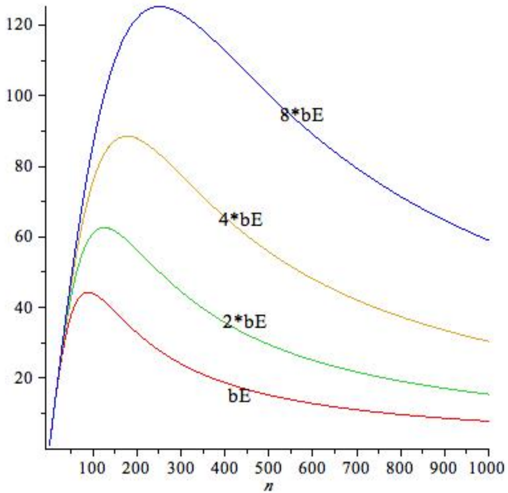
bwGRiD – Two Regions

Speed-up comparison of measurements and model for two regions for an estimated bandwidth contention of $p(n) = n/20$



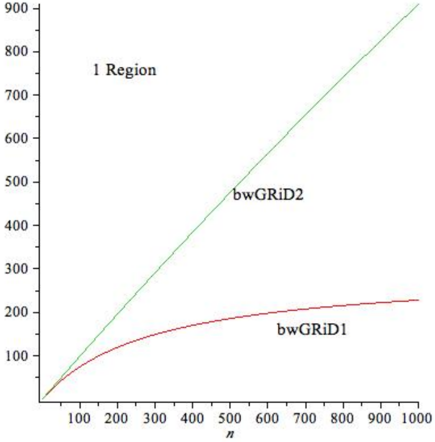
bwGRiD – Two Regions

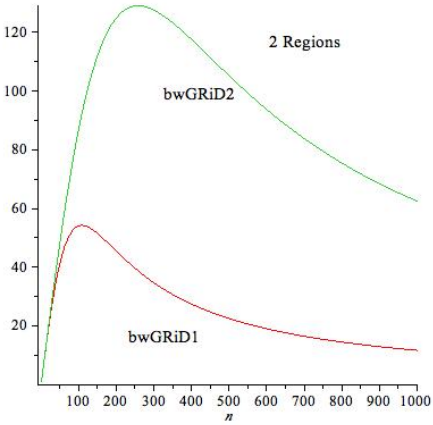
Speed-up bwGRiD1 for two regions and varying bandwidth B^E.



bwGRiD – Speedup prediction

Speed-up in bwGRiD1&2 for one and two regions with $n_p = 40000$.





Conclusions

- Performance model is based on roofline model
- Throughput and speed-up are described by 2 – 3 scaling parameters which depend on important hardware and software characterisitics
- Model reproduces LinPack measurements for one and two regions (bwGRiD1)
- Model predicts performance of next generation system (bwGRiD2)
- Upper bounds for region sizes are derived by analyzing the maximal Speedup
- Lower bounds for region sizes are derived by analyzing the $n_{1/2}$ values (see paper)
- Next steps:
  - More detailed model for the communication within a region
  - Investigation of other applications