Amdahl’s Laws and Extreme Data-Intensive Scientific Computing

Alex Szalay
The Johns Hopkins University
Scientific Data Analysis Today

- Scientific data is doubling every year, reaching PBs
- Data is everywhere, never will be at a single location
- Need randomized, incremental algorithms
  - *Best result in 1 min, 1 hour, 1 day, 1 week*
- Architectures increasingly CPU-heavy, IO-poor
- Data-intensive scalable architectures needed
- Most scientific data analysis done on small to midsize BeoWulf clusters, from faculty startup
- Universities hitting the “power wall”
- Soon we cannot even store the incoming data stream
- **Not scalable, not maintainable…**
Continuing Growth of Data

How long does the growth continue?
• High end always linear
• Exponential comes from technology + economics
  – rapidly changing generations of sensors and processing
  – like CCD’s replacing photographic plates
• How many generations of instruments are left?
• Are there new growth areas emerging?
• **Software is becoming a new kind of instrument**
  – Value added federated data sets
  – Large and complex simulations
  – Hierarchical data replication
Challenges

• Where are the challenges in DISC today?
  – Storage size
  – System balance
  – Data mobility
  – Statistical algorithms
  – Scalability/power

• What is being done to soften it?
  – Scale up or scale out…
  – New SW platforms emerging
  – Testing disruptive technologies
  – New algorithms
Increased Diversification

One shoe does not fit all!

• Diversity grows naturally, no matter what
• Evolutionary pressures help
  – Large floating point calculations move to GPUs
  – Fast IO moves to high Amdahl number systems
  – Stream processing emerging
  – noSQL vs databases vs column store etc
  – Large data moves into the cloud (but how?)

• Individual groups want subtle specializations

At the same time

• What remains in the middle/mainstream?
• Boutique systems dead, commodity rules
• Large graph problems still hard to do (XMT or Pregel)
Amdahl’s Laws

Gene Amdahl (1965): Laws for a balanced system

i. Parallelism: max speedup is $S/(S+P)$

ii. One bit of IO/sec per instruction/sec (BW)

iii. One byte of memory per one instruction/sec (MEM)

Modern multi-core systems move farther away from Amdahl’s Laws
(Bell, Gray and Szalay 2006)
## Typical Amdahl Numbers

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RAM</td>
</tr>
<tr>
<td>BeoWulf</td>
<td>100</td>
<td>300</td>
<td>200</td>
<td>3000</td>
<td>0.67</td>
</tr>
<tr>
<td>Desktop</td>
<td>2</td>
<td>6</td>
<td>4</td>
<td>150</td>
<td>0.67</td>
</tr>
<tr>
<td>Cloud VM</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>30</td>
<td>1.33</td>
</tr>
<tr>
<td>SC1</td>
<td>212992</td>
<td>150000</td>
<td>18600</td>
<td>16900</td>
<td>0.12</td>
</tr>
<tr>
<td>SC2</td>
<td>2090</td>
<td>5000</td>
<td>8260</td>
<td>4700</td>
<td>1.65</td>
</tr>
<tr>
<td>GrayWulf</td>
<td>416</td>
<td>1107</td>
<td>1152</td>
<td>70000</td>
<td>1.04</td>
</tr>
</tbody>
</table>
The Data Sizes Involved
DISC Needs Today

• Disk space, disk space, disk space!!!!
• Current problems not on Google scale yet:
  – 10-30TB easy, 100TB doable, 300TB really hard
  – For detailed analysis we need to park data for several months
• Sequential IO bandwidth
  – If not sequential for large data set, we cannot do it
• Stream processing
  – A continuum of Amdahl numbers
• How do can move 100TB within a University?
  – 1Gbps 10 days
  – 10 Gbps 1 day (but need to share backbone)
  – 100 lbs box few hours
• From outside?
  – Dedicated 10Gbps or FedEx
Tradeoffs Today

Stu Feldman: Extreme computing is about tradeoffs

Ordered priorities for data-intensive scientific computing

1. Total storage  (-> low redundancy)
2. Cost  (-> total cost vs price of raw disks)
3. Sequential IO  (-> locally attached disks, fast ctrl)
4. Fast stream processing  (->GPUs inside server)
5. Low power  (-> slow normal CPUs, lots of disks/mobo)

The order will be different in a few years...and scalability may appear as well
## Cost of a Petabyte

**COST OF A PETABYTE**

<table>
<thead>
<tr>
<th>Manufacturer</th>
<th>Model</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RAW DRIVES</strong></td>
<td></td>
<td>$81,000</td>
</tr>
<tr>
<td><strong>BACKBLAZE</strong></td>
<td></td>
<td>$117,000</td>
</tr>
<tr>
<td><strong>DELL</strong></td>
<td>MD1000</td>
<td>$826,000</td>
</tr>
<tr>
<td><strong>Sun Microsystems</strong></td>
<td>X4550</td>
<td>$1,000,000</td>
</tr>
<tr>
<td><strong>NetApp</strong></td>
<td>FAS-6000</td>
<td>$1,714,000</td>
</tr>
<tr>
<td><strong>AMAZON S3</strong></td>
<td></td>
<td>$2,806,000</td>
</tr>
<tr>
<td><strong>EMC</strong></td>
<td>EMC NS-960</td>
<td>$2,860,000</td>
</tr>
</tbody>
</table>

*Amazon S3 Storage over three years (minus electricity, co-location and administration).*
JHU Data-Scope

- Funded by NSF MRI to build a new ‘instrument’ to look at data
- Goal: 100 servers for $1.2M, including 10G switches
- Two-tier: performance (P) and storage (S)
- Large (5PB) + cheap + fast (400+GBps), but …
  . . . . a special purpose instrument
**Proposed Projects at JHU**

<table>
<thead>
<tr>
<th>Discipline</th>
<th>data [TB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Astrophysics</td>
<td>930</td>
</tr>
<tr>
<td>HEP/Material Sci.</td>
<td>394</td>
</tr>
<tr>
<td>CFD</td>
<td>425</td>
</tr>
<tr>
<td>BioInformatics</td>
<td>414</td>
</tr>
<tr>
<td>Environmental</td>
<td>660</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>2823</strong></td>
</tr>
</tbody>
</table>

19 projects total proposed for the Data-Scope, more coming, data lifetimes between 3 mo and 3 yrs
Solid State Disks (SSDs)

- Much higher throughput with lower power consumption
  - 250MB/s sequential read, 200MB/s sequential write
  - 1-2W peak power
- Incorporating SSDs to server designs
  - *Scale-up*
    - Install SSDs into existing high end servers
    - Quickly running out of PCI BW
  - *Scale down*
    - Amdahl Blade: one SSD per core
    - Ideal Amdahl number and IOPS ratio
Cyberbricks/Amdahl Blades

- **Scale down** the CPUs to the disks!
  - *Solid State Disks (SSDs)*
  - *1 low power CPU per SSD*
- **Current SSD parameters**
  - *OCZ Vertex 120GB, 250MB/s read, 10,000+ IOPS, $350*
  - *Power consumption 0.2W idle, 1-2W under load*
- **Low power motherboards**
  - *Intel dual Atom N330 + NVIDIA ION chipset 28W at 1.6GHz*
- **Combination achieves perfect Amdahl blade**
  - *200MB/s=1.6Gbits/s ⇔ 1.6GHz of Atom*
Building a Low Power Cluster

*AS, G.Bell, H.Huang, A.Terzis, A.White (HotPower09):*

*Evaluation of many different motherboard + SSD combinations*

<table>
<thead>
<tr>
<th>System</th>
<th>Model</th>
<th>CPU</th>
<th>Chipset</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASUS</td>
<td>EeeBox</td>
<td>N270</td>
<td>945GSE</td>
</tr>
<tr>
<td>Intel</td>
<td>D945GCLF2</td>
<td>N330</td>
<td>945GC</td>
</tr>
<tr>
<td>Zotac</td>
<td>Ion</td>
<td>N330</td>
<td>ION</td>
</tr>
<tr>
<td>AxiomTek</td>
<td>Pico 820</td>
<td>Z530</td>
<td>US15W</td>
</tr>
<tr>
<td>Alix</td>
<td>3C2</td>
<td>LX800</td>
<td>AMD</td>
</tr>
</tbody>
</table>

*Sweet spot: dual core Atom!*
Cost of ownership includes $0.15/kWh x 1.6 for AC overhead
Price assumes 250GB SSD for $400

1GB/sec sequential IO for $2400
1kW Amdahl Cluster

• 36-node Amdahl cluster using 1000W total
  – Zotac Atom/ION motherboards
  – 4GB of memory, N330 dual core Atom, 16 GPU cores

• Aggregate disk space 43.6TB
  – 63 x 120GB SSD = 7.7 TB
  – 27x 1TB Samsung F1 = 27.0 TB
  – 18x.5TB Samsung M1= 9.0 TB

• I/O Performance: 18GB/s

• Amdahl number = 1 for under $40K

• Using the GPUs for data mining:
  – 6.4B multidimensional regressions in 5 minutes over 1.2TB
  – Ported RF module from R in C#/CUDA
Histogram Test

- Build histogram of 544 million objects from 1.2TB
- Reasonably complex selection criteria
- Distributed SQL query with dynamic load balancing
- Runs in 100 sec, average throughput: **13GB/sec**
- SSD at 421MB/s, SSD+1TB at 397MB/s & 379 MB/s
## Scaling I

<table>
<thead>
<tr>
<th>CPU</th>
<th>seq IO</th>
<th>randIO</th>
<th>disk</th>
<th>power</th>
<th>cost</th>
<th>rel power</th>
<th>price</th>
<th>elect</th>
</tr>
</thead>
<tbody>
<tr>
<td>GrayWulf</td>
<td>21.3</td>
<td>1.5</td>
<td>6</td>
<td>22.5</td>
<td>1150</td>
<td>19253</td>
<td>1.0000</td>
<td>12000.0</td>
</tr>
<tr>
<td>ASUS</td>
<td>37.6</td>
<td>2.9</td>
<td>108</td>
<td>5.9</td>
<td>446</td>
<td>19253</td>
<td>0.3880</td>
<td>16439.0</td>
</tr>
<tr>
<td>Intel</td>
<td>52.4</td>
<td>8.2</td>
<td>164</td>
<td>8.2</td>
<td>458</td>
<td>19253</td>
<td>0.3984</td>
<td>16363.5</td>
</tr>
<tr>
<td>Zotac</td>
<td>51.8</td>
<td>8.1</td>
<td>168</td>
<td>8.1</td>
<td>486</td>
<td>19253</td>
<td>0.4223</td>
<td>16189.9</td>
</tr>
<tr>
<td>Pico820</td>
<td>31.0</td>
<td>2.3</td>
<td>77</td>
<td>4.8</td>
<td>290</td>
<td>19253</td>
<td>0.2525</td>
<td>17421.9</td>
</tr>
<tr>
<td>Alix 3C2</td>
<td>42.7</td>
<td>2.1</td>
<td>0.7</td>
<td>342</td>
<td>9917</td>
<td>0.1999</td>
<td>8467.7</td>
<td>1449.6</td>
</tr>
<tr>
<td>hybrid</td>
<td>56.8</td>
<td>8.9</td>
<td>40.0</td>
<td>799</td>
<td>19253</td>
<td>0.6951</td>
<td>14211.4</td>
<td>5041.9</td>
</tr>
</tbody>
</table>

### Sweet spot

<table>
<thead>
<tr>
<th>CPU</th>
<th>seq IO</th>
<th>randIO</th>
<th>disk</th>
<th>power</th>
<th>cost</th>
<th>rel power</th>
<th>price</th>
<th>elect</th>
</tr>
</thead>
<tbody>
<tr>
<td>GrayWulf</td>
<td>21.3</td>
<td>1.5</td>
<td>6</td>
<td>22.5</td>
<td>1150</td>
<td>19253</td>
<td>1.0000</td>
<td>12000.0</td>
</tr>
<tr>
<td>ASUS</td>
<td>19.4</td>
<td>1.5</td>
<td>56</td>
<td>3.0</td>
<td>230</td>
<td>9917</td>
<td>0.1999</td>
<td>8467.7</td>
</tr>
<tr>
<td>Intel</td>
<td>9.6</td>
<td>1.5</td>
<td>30</td>
<td>1.5</td>
<td>84</td>
<td>3530</td>
<td>0.0730</td>
<td>3000.0</td>
</tr>
<tr>
<td>Zotac</td>
<td>9.6</td>
<td>1.5</td>
<td>31</td>
<td>1.5</td>
<td>90</td>
<td>3568</td>
<td>0.0783</td>
<td>3000.0</td>
</tr>
<tr>
<td>Pico820</td>
<td>20.0</td>
<td>1.5</td>
<td>50</td>
<td>3.1</td>
<td>188</td>
<td>12433</td>
<td>0.1630</td>
<td>11250.0</td>
</tr>
<tr>
<td>Alix 3C2</td>
<td>30.0</td>
<td>1.5</td>
<td>0</td>
<td>0.5</td>
<td>240</td>
<td>13514</td>
<td>0.2087</td>
<td>12000.0</td>
</tr>
<tr>
<td>hybrid</td>
<td>9.6</td>
<td>1.5</td>
<td>18</td>
<td>6.8</td>
<td>135</td>
<td>3251</td>
<td>0.1174</td>
<td>2400.0</td>
</tr>
</tbody>
</table>
As soon as we have a 750GB SSD for $400, it will be cheaper to build PB-scale storage systems from Amdahl blades!!!
Arrays in SQL Server

- Recent effort by Laszlo Dobos
- Written in C++
- Arrays packed into varbinary(8000) or varbinary(max)
- Various subsets, aggregates, extractions and conversions in T-SQL (see regrid example:)

```sql
SELECT s.ix, DoubleArray.Avg(s.a)
INTO ##temptable
FROM DoubleArray.Split(@a, Int16Array.Vector_3(4,4,4)) s
--
SELECT @subsample = DoubleArray.Concat_N('##temptable')
--
@a is an array of doubles with 3 indices
The first command averages the array over 4×4×4 blocks,
returns indices and the value of the average into a table
Then we build a new (collapsed) array from its output
```
GPUs

- Hundreds of cores – 100K+ parallel threads!
- CPU is free, RAM is slow
- Amdahl’s first Law applies
- GPU has >50GB/s bandwidth inside
- Still difficult to keep the cores busy
  - IO bandwidth limiting many computations

- How to integrate it with data intensive computations?
- How to integrate it with SQL?

See Budavari talk
Short Term Trends in DISC

• Large data sets are here, solutions are not
  – 100TB is the current practical limit in science projects
  – Much wider need for DISC than expected
  – Next year the pain threshold will be about 500TB
• No real data-intensive computing facilities available
• Even HPC projects choking on IO, and lack storage
• Cloud computing tradeoffs different from science needs
• Increased diversification using commodity components
  – SSDs, GPUs (Gordon, Amdahl blades,…)
• Scientists are “frugal”, also pushing the limit
  – Exploring disruptive technologies…
  – Still building our own…on the cheap
  – Finally, campus level aggregation emerging …