High Performance Computing, Cloud computing, and Remote Sensing Big Data

Lizhe Wang

Institute of Remote Sensing and Digital Earth (RADI)
Chinese Academy of Sciences (CAS)
email: Lizhe.Wang@computer.org

March 2016, Canberra, Australia
1. RS Big Data: an Overview

2. How Big? a Quantitative Model

3. Software Systems for High Performance RS Big Data Processing
   - Global Distributed RS Data Model in Memory
   - Generic Parallel Processing Model for RS Big Data
   - Parallel Storage Model for RS Big Data
   - Scheduling of Large-scale RS images Parallel Processing

4. RS Big Data Processing in Clouds: Current Work

5. Concluding Remark
1. RS Big Data: an Overview

2. How Big? a Quantitative Model

3. Software Systems for High Performance RS Big Data Processing
   - Global Distributed RS Data Model in Memory
   - Generic Parallel Processing Model for RS Big Data
   - Parallel Storage Model for RS Big Data
   - Scheduling of Large-scale RS images Parallel Processing

4. RS Big Data Processing in Clouds: Current Work

5. Concluding Remark
- Nimbus, 1964
- 514 EO satellites till Dec. 2011
- EO has played an important role for human being society
EO satellites in 1962-2012
Plan of EO satellites in 2013-2035
China Remote Sensing Satellite Ground Station, a member of the Landsat Ground Station Operations Working Group, boasts one of the world’s highest capacities for receiving, processing, and distributing satellite data. With over 3.3 million scenes of satellite data accumulated on file since 1986, it is regarded as the largest Earth observation satellite data archive in China.

Its three stations at Miyun, Kashi, and Sanya can receive data simultaneously from satellites covering the whole territory of China and 70% of Asia.

It receives 12 satellites currently, will receive 50 satellite in 2025, and become the LARGEST ground station for civil use in the world.
<table>
<thead>
<tr>
<th>编号</th>
<th>卫星名称</th>
<th>所属国家或组织</th>
<th>开始接收时间</th>
<th>目前运行状况</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LANDSAT-5</td>
<td>美国</td>
<td>1986</td>
<td>✓</td>
</tr>
<tr>
<td>2</td>
<td>ERS-1</td>
<td>欧空局</td>
<td>1993</td>
<td>✓</td>
</tr>
<tr>
<td>3</td>
<td>JERS-1</td>
<td>日本</td>
<td>1993</td>
<td>✓</td>
</tr>
<tr>
<td>4</td>
<td>ERS-2</td>
<td>欧空局</td>
<td>1995</td>
<td>✓</td>
</tr>
<tr>
<td>5</td>
<td>RADARSAT-1</td>
<td>加拿大</td>
<td>1997</td>
<td>✓</td>
</tr>
<tr>
<td>6</td>
<td>SPOT-1</td>
<td>法国</td>
<td>1998</td>
<td>✓</td>
</tr>
<tr>
<td>7</td>
<td>SPOT-2</td>
<td>法国</td>
<td>1998</td>
<td>✓</td>
</tr>
<tr>
<td>8</td>
<td>CBERS-01</td>
<td>中国</td>
<td>1999</td>
<td>✓</td>
</tr>
<tr>
<td>9</td>
<td>SPOT-4</td>
<td>法国</td>
<td>1999</td>
<td>✓</td>
</tr>
<tr>
<td>10</td>
<td>LANDSAT-7</td>
<td>美国</td>
<td>2000</td>
<td>✓</td>
</tr>
<tr>
<td>11</td>
<td>SPOT-5</td>
<td>法国</td>
<td>2002</td>
<td>✓</td>
</tr>
<tr>
<td>12</td>
<td>ENVISAT</td>
<td>欧空局</td>
<td>2003</td>
<td>✓</td>
</tr>
<tr>
<td>13</td>
<td>CBERS-02</td>
<td>中国</td>
<td>2003</td>
<td>✓</td>
</tr>
<tr>
<td>14</td>
<td>RESOURCESAT-1</td>
<td>印度</td>
<td>2005</td>
<td>✓</td>
</tr>
<tr>
<td>15</td>
<td>CBERS-02B</td>
<td>中国</td>
<td>2007</td>
<td>✓</td>
</tr>
<tr>
<td>16</td>
<td>RADARSAT-2</td>
<td>加拿大</td>
<td>2008</td>
<td>✓</td>
</tr>
<tr>
<td>17</td>
<td>HJ-1A</td>
<td>中国</td>
<td>2008</td>
<td>✓</td>
</tr>
<tr>
<td>18</td>
<td>HJ-1B</td>
<td>中国</td>
<td>2008</td>
<td>✓</td>
</tr>
<tr>
<td>19</td>
<td>THEOS</td>
<td>泰国</td>
<td>2011</td>
<td>✓</td>
</tr>
<tr>
<td>20</td>
<td>资源一号02C</td>
<td>中国</td>
<td>2011</td>
<td>✓</td>
</tr>
<tr>
<td>21</td>
<td>资源三号</td>
<td>中国</td>
<td>2012</td>
<td>✓</td>
</tr>
<tr>
<td>22</td>
<td>实践九号A星</td>
<td>中国</td>
<td>2012</td>
<td>✓</td>
</tr>
<tr>
<td>23</td>
<td>实践九号B星</td>
<td>中国</td>
<td>2012</td>
<td>✓</td>
</tr>
<tr>
<td>24</td>
<td>HJ-1C</td>
<td>中国</td>
<td>2012</td>
<td>✓</td>
</tr>
</tbody>
</table>
1. RS Big Data: an Overview

2. How Big? a Quantitative Model

3. Software Systems for High Performance RS Big Data Processing
   - Global Distributed RS Data Model in Memory
   - Generic Parallel Processing Model for RS Big Data
   - Parallel Storage Model for RS Big Data
   - Scheduling of Large-scale RS images Parallel Processing

4. RS Big Data Processing in Clouds: Current Work

5. Concluding Remark
### Data acquisition: data volumes and data rates

<table>
<thead>
<tr>
<th>Satellites</th>
<th>Data Rates (MB/S)</th>
<th>Data Amount (GB/Day)</th>
<th>Data Amount (TB/Year)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZY-02</td>
<td>113.98</td>
<td>115</td>
<td>40.99</td>
</tr>
<tr>
<td>ZY-03</td>
<td>304</td>
<td>334</td>
<td>119.05</td>
</tr>
<tr>
<td>HJ-1A</td>
<td>120</td>
<td>114</td>
<td>40.63</td>
</tr>
<tr>
<td>HJ-1B</td>
<td>60</td>
<td>57</td>
<td>20.32</td>
</tr>
<tr>
<td>HJ-1C</td>
<td>320</td>
<td>187.5</td>
<td>66.83</td>
</tr>
<tr>
<td>LandSat8</td>
<td>440.00</td>
<td>241.70</td>
<td>47.33</td>
</tr>
<tr>
<td>RadaSat2</td>
<td>105.00</td>
<td>57.68</td>
<td>11.29</td>
</tr>
<tr>
<td>RadaSat-1</td>
<td>105.00</td>
<td>57.68</td>
<td>11.29</td>
</tr>
<tr>
<td>SPOT-5</td>
<td>100.00</td>
<td>54.93</td>
<td>10.76</td>
</tr>
<tr>
<td>LANDSAT5</td>
<td>85.00</td>
<td>28.02</td>
<td>9.99</td>
</tr>
<tr>
<td>SPOT-4</td>
<td>50.00</td>
<td>10.99</td>
<td>3.92</td>
</tr>
<tr>
<td>RADARSAT-1</td>
<td>105.00</td>
<td>34.61</td>
<td>12.34</td>
</tr>
<tr>
<td>ENVISAT</td>
<td>100.00</td>
<td>32.96</td>
<td>11.75</td>
</tr>
<tr>
<td>IRS-P6</td>
<td>210.00</td>
<td>46.14</td>
<td>16.45</td>
</tr>
</tbody>
</table>

- Single RS data set: GB
- Daily receiving RS data set: TB
- National RS data archive: PB
- International RS data archive: EB
## Data volumes of a single scene of L0, L1, and L2 Products

<table>
<thead>
<tr>
<th>Satellites(Sensor)</th>
<th>L0(MB)</th>
<th>L1(MB)</th>
<th>L2(MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BJ-1(MS)</td>
<td>228.3</td>
<td>228.3</td>
<td>333.32</td>
</tr>
<tr>
<td>BJ-1(PAN)</td>
<td>98.9</td>
<td>98.9</td>
<td>150.33</td>
</tr>
<tr>
<td>SPOT-5(PAN)</td>
<td>572</td>
<td>572</td>
<td>45.96</td>
</tr>
<tr>
<td>SPOT-5(MS)</td>
<td>144</td>
<td>144</td>
<td>252.21</td>
</tr>
<tr>
<td>LANDSAT7(ETM)</td>
<td>337</td>
<td>337</td>
<td>532</td>
</tr>
<tr>
<td>LANDSAT8(UTM)</td>
<td>481</td>
<td>481</td>
<td>781</td>
</tr>
</tbody>
</table>
## Data volume of value-added processing

<table>
<thead>
<tr>
<th>Satellites (Sensor)</th>
<th>Fine 1 Scene (MB)</th>
<th>Ortho 1 Scene (MB)</th>
<th>Regional Mosaic (GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LANDSAT7</td>
<td>511.52</td>
<td>515.3</td>
<td>7.48 255.5</td>
</tr>
<tr>
<td>ZY-02C (PAN)</td>
<td>201.61</td>
<td>206.65</td>
<td>N/A N/A</td>
</tr>
<tr>
<td>ZY-02C (HRI)</td>
<td>467.3</td>
<td>475.15</td>
<td>N/A N/A</td>
</tr>
</tbody>
</table>
Data Volume Throughout the Entire Flow

Data Amount (Log(D)): Log(MB)

CBERS-02C(HRI)  LANDSAT5
Comparative analysis of the data throughput rate
## Computational Complexity of Algorithms

<table>
<thead>
<tr>
<th>Stage</th>
<th>Algorithm</th>
<th>Complexity</th>
<th>Parallelism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preprocessing</td>
<td>Radiometric</td>
<td>$O(n)$</td>
<td>Excellent</td>
</tr>
<tr>
<td></td>
<td>Geometric</td>
<td>$O(n^2)$</td>
<td>Excellent</td>
</tr>
<tr>
<td></td>
<td>Registration</td>
<td>$O(n^2 \log(n))$</td>
<td>Excellent</td>
</tr>
<tr>
<td>Value-added Processing</td>
<td>K-T Trans</td>
<td>$O(p^2 bn^2)$</td>
<td>Good</td>
</tr>
<tr>
<td></td>
<td>Convolution</td>
<td>$O(k^2 n^2)$</td>
<td>Good</td>
</tr>
<tr>
<td></td>
<td>Mosaicking</td>
<td>$O(n^3)$</td>
<td>Normal</td>
</tr>
<tr>
<td>Information Extraction</td>
<td>K-mean</td>
<td>$O(rbmn^2)$</td>
<td>Normal</td>
</tr>
<tr>
<td></td>
<td>Bayers</td>
<td>$O(bmn^2)$</td>
<td>good</td>
</tr>
<tr>
<td></td>
<td>BP</td>
<td>$O(n^2 \log(n))$</td>
<td>good</td>
</tr>
</tbody>
</table>
Computational Complexity of Algorithms

Lizhe Wang  HPC, Clouds, RS Big Data
The $D_{IR_S}$ model

Choose Key Factors

DI Experiments on Typical Algorithms (Flops/IO, Throughput, Amount)

Computational Bandwidth (When Balanced)

Empirical Estimation of DI Index (Empirical DI Curves)

Modeling DI Indicator
$DI = F(C,D,N,T..)$

Choose Typical Algorithm

Estimation of DI
The $D_{IR_{\text{S}}} \text{ model}$

\[
D_{IR_{\text{S}}} = \begin{cases} 
0 & \text{if } \frac{10^{DC}}{60^T} < 1 \\
\log \frac{10^{DC}}{60^T} & \text{if } \frac{10^{DC}}{60^T} > 1
\end{cases}
\]
$D = \max \{ D_{in}, D_{m1}, D_{m2}, \ldots, D_{out} \}$ (2)
Table: Quantization of computational complexity ‘C’

<table>
<thead>
<tr>
<th>Computational Complexity</th>
<th>‘C’</th>
</tr>
</thead>
<tbody>
<tr>
<td>$O(n)$</td>
<td>1</td>
</tr>
<tr>
<td>$O(n^2)$</td>
<td>2</td>
</tr>
<tr>
<td>$O(n^2 \log(n))$</td>
<td>2.5</td>
</tr>
<tr>
<td>$O(k^2 n^2)$</td>
<td>2.2</td>
</tr>
<tr>
<td>$O(n^3)$</td>
<td>3</td>
</tr>
<tr>
<td>$O(n^3) \sim O(n^4)$</td>
<td>4</td>
</tr>
</tbody>
</table>
Table: Quantization of Time Requirement ‘T’

<table>
<thead>
<tr>
<th>Time Requirement</th>
<th>‘T’</th>
</tr>
</thead>
<tbody>
<tr>
<td>Second</td>
<td>1</td>
</tr>
<tr>
<td>Minute</td>
<td>2</td>
</tr>
<tr>
<td>Hour</td>
<td>3</td>
</tr>
<tr>
<td>Day</td>
<td>3.5</td>
</tr>
<tr>
<td>Week</td>
<td>4</td>
</tr>
<tr>
<td>&gt;Month</td>
<td>5</td>
</tr>
</tbody>
</table>
Empirical Estimation of $D_{IRS}$ Through the Entire RS Data Processing Flow

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>‘C’</th>
<th>‘D’</th>
<th>‘T’</th>
<th>$D_{IRS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>L2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2.4</td>
</tr>
<tr>
<td>L3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2.4</td>
</tr>
<tr>
<td>L4</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2.4</td>
</tr>
<tr>
<td>MO</td>
<td>5.5</td>
<td>5.5</td>
<td>3</td>
<td>11.2</td>
</tr>
<tr>
<td>CL</td>
<td>5.5</td>
<td>3</td>
<td>3</td>
<td>8.4</td>
</tr>
<tr>
<td>AR</td>
<td>5.5</td>
<td>3</td>
<td>3</td>
<td>11.2</td>
</tr>
<tr>
<td>OP</td>
<td>4.5</td>
<td>3</td>
<td>3</td>
<td>3.7</td>
</tr>
</tbody>
</table>
Empirical Estimation of the $D_{IR}$ Indicator

Lizhe Wang  
HPC, Clouds, RS Big Data
Some Findings for RS Big Data

- Big Data
- Compute-intensive and data-intensive
- Unbalanced Data Flow
- Complex Algorithm Flow
1 RS Big Data: an Overview

2 How Big? a Quantitative Model

3 Software Systems for High Performance RS Big Data Processing
   - Global Distributed RS Data Model in Memory
   - Generic Parallel Processing Model for RS Big Data
   - Parallel Storage Model for RS Big Data
   - Scheduling of Large-scale RS images Parallel Processing

4 RS Big Data Processing in Clouds: Current Work

5 Concluding Remark
Philosophy of RS Big Data Computing

- Data processing models and algorithms
  - Approximate computation
  - Reduced Computation
  - Partial Computation

- System and Software
  - Platforms: Clusters, Data Centers, Clouds, High-speed networking
  - Software system: Parallel file systems, High Performance I/O, Programming models, Memory models, Runtime System

- Paradigm of RS Data Product Service
  - Precise Prediction, On-Demand Computing, Active Service
  - Standalization, Serialisation and Automation of RS Data production
Philosophy of RS Big Data Computing

- Data processing models and algorithms
  - Approximate computation
  - Reduced Computation
  - Partial Computation

- System and Software
  - Platforms: Clusters, Data Centers, Clouds, High-speed networking
  - Software system: Parallel file systems, High Performance IO, Programming models, Memory models, Runtime Systems

- Paradigm of RS Data Product Service
  - Precise Prediction, On-Demand Computing, Active Service
  - Standalization, Serialisation and Automation of RS Data production
High Performance Processing for EO Big Data

Where:
- EO Big Data and its Applications

What:
- Quantitative analysis of EO Big Data and its processing algorithms
- Structured, semi-structured and unstructured EO Big Data
- High performance processing algorithms and software models

Systems and Platforms
- Clusters
- Satellite data centres
- Clouds, and Multiple datacenters

Data Model
- Memory model: GDDM
- Storage model: HPGFS
- IO model: ADIOS, HA

Processing Model
- Distributed Memory Processing Model (MPI)
- Workflow Model
- Batch Processing Model
- Streaming data

Production System
- Parallel RS Image Processing System (PIPS, PIPS+)
- RS Data Production System across Multiple Satellite Datacenters
Contents

1 RS Big Data: an Overview

2 How Big? a Quantitative Model

3 Software Systems for High Performance RS Big Data Processing
   • Global Distributed RS Data Model in Memory
   • Generic Parallel Processing Model for RS Big Data
   • Parallel Storage Model for RS Big Data
   • Scheduling of Large-scale RS images Parallel Processing

4 RS Big Data Processing in Clouds: Current Work

5 Concluding Remark
A System View

Hilbert curve

HPC, Clouds, RS Big Data
Contents

1 RS Big Data: an Overview

2 How Big? a Quantitative Model

3 Software Systems for High Performance RS Big Data Processing
   • Global Distributed RS Data Model in Memory
   • Generic Parallel Processing Model for RS Big Data
   • Parallel Storage Model for RS Big Data
   • Scheduling of Large-scale RS images Parallel Processing

4 RS Big Data Processing in Clouds: Current Work

5 Concluding Remark
Generic Parallel Processing Model for RS Big Data

RS-GPPS

Template
<Class T>
Class
Skeleton{}

Generic
Algorithm
Skeletons
(Templates)

Distributed
RS Data
Templates

Data Distribution
Info

MPI Runtime Environment

Multicore Cluster

Parallel FS

data block load/write

Implement Distributed RS Data
1) Slice RS data logically into overlapped blocks;
2) Distributed blocks across nodes;
3) Converged blocks as a Global RS data via one-sided messaging of MPI

Implement skeleton
(Perform Computations on Distributed RS Data)
1) Divide the Task into subtasks;
2) Load the Data Blocks concurrently through parallel I/O;
3) Execute Sequential Codes in parallel
1. RS Big Data: an Overview

2. How Big? a Quantitative Model

3. Software Systems for High Performance RS Big Data Processing
   - Global Distributed RS Data Model in Memory
   - Generic Parallel Processing Model for RS Big Data
   - Parallel Storage Model for RS Big Data
   - Scheduling of Large-scale RS images Parallel Processing

4. RS Big Data Processing in Clouds: Current Work

5. Concluding Remark
RS Data Access Pattern

(a) Neighbor-Based Processing (Regional dependent)  
(b) Rectangular-Block Access Pattern
(a) Band-Related Processing (Band dependent)

(b) Cross-File Access Pattern
RS Data Access Pattern

(a) Irregular Processing (Global or Irregular Dependent)

(b) Diagonal Access Pattern

(c) Polygon Access Pattern

Lizhe Wang  HPC, Clouds, RS Big Data
Optimized Data Layout in HPGFS

Distributed RS Data Object

Geographical Metadata
- Image Info
- Map Info
- Projection Para
- Satellite & Sensor Info

Image Data
- Image Band 1
- Image Band 2
- Image Band 3
- Image Band n

Basic Operations
- getGeoMetadata()
- setGeoMetadata()
- projectionReform()
- resampling()

Geographical Metadata
- Image Info
- Map Info
- Projection Para (differ with projections)

Satellite & Sensor Para (differ with sensors)
- UTM
  - Ellipsoid Datum
  - UTM Zone
- Transvers Mercator
  - Ellipsoid Datum
  - Central meridian
  - Coordinating origin
- ETM+
  - Sensor type
  - Orbit
  - Semi-major
  - Semi-minor
  - Orbital inclination
- Radarsat
  - Sensor type
  - Orbit
  - Wave length
  - Range Resolution
  - Azimuth

Metadata Server

Geodata

Metadata

Lizhe Wang
HPC, Clouds, RS Big Data
*Lizhe Wang, Yan Ma, etc. A Parallel File System with Application-aware Data Layout Policies in Digital Earth*, IEEE Transactions on Parallel & Distributed Systems, doi:10.1109/TPDS.2014.2322362
Contents

1 RS Big Data: an Overview

2 How Big? a Quantitative Model

3 Software Systems for High Performance RS Big Data Processing
   - Global Distributed RS Data Model in Memory
   - Generic Parallel Processing Model for RS Big Data
   - Parallel Storage Model for RS Big Data
   - Scheduling of Large-scale RS images Parallel Processing

4 RS Big Data Processing in Clouds: Current Work

5 Concluding Remark
Input file; XML file; Individual Executable Program; Output file;

Individual computing task

Task tree based parallel computing

Original sequence

Lizhe Wang
HPC, Clouds, RS Big Data
Large-scale RS images Parallel Processing

Master Node
Calculating Overlapping Region

Task partition

Slave Nodes (parallel)
Re-Projection
Registration
(maximum mutual information)
Abstract Control Point
Geometric Transformation

Radiometric Balancing
Histogram Matching

Resampling
Seam Line Selection
Blending

Image Arrangement
Row-major sorting
Path-major sorting

Image Sequence
1, 11, 12, 15, 0, 7, 1, 8, 5, 2, 4, 10, 3, 5, 14, 6

arranged images

Task Tree Creating
Generating Adjacent Relations
Task Partition

"messy" images
geometric metadata

Lizhe Wang
HPC, Clouds, RS Big Data

Lizhe Wang

HPC, Clouds, RS Big Data
Data: More than Half of China. (20° N–47° N and 84° E–180° E) 352 Scenes of ETM+, 15m resolution, PAN

Performance: On a 10-node cluster, it takes 5 hours. typically it is several-day work with a commercial software
1. RS Big Data: an Overview

2. How Big? a Quantitative Model

3. Software Systems for High Performance RS Big Data Processing
   - Global Distributed RS Data Model in Memory
   - Generic Parallel Processing Model for RS Big Data
   - Parallel Storage Model for RS Big Data
   - Scheduling of Large-scale RS images Parallel Processing

4. RS Big Data Processing in Clouds: Current Work

5. Concluding Remark
Multiple Datacenter Collaborative Process System (MCCPS)

- Join 6 RS data centers of China;
- Process 12+ multi-source RS data;
- The amount of data is over 1 PB

<table>
<thead>
<tr>
<th>Data Center</th>
<th>Satellite</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. CRESDA</td>
<td>CBERS, HJ, ZY</td>
</tr>
<tr>
<td>2. NSOAS</td>
<td>HY</td>
</tr>
<tr>
<td>3. NSMC</td>
<td>FY3A, FY3B, FY2E</td>
</tr>
<tr>
<td>4. CNIC</td>
<td>LandSat, MODIS</td>
</tr>
<tr>
<td>5. 21AT</td>
<td>BJ-1</td>
</tr>
<tr>
<td>6. RADI</td>
<td>ENVISAT-1, IRS-P6</td>
</tr>
<tr>
<td>Data Volume</td>
<td>&gt;1PB</td>
</tr>
</tbody>
</table>
Multi-datacenter Clouds for RS Data Processing

- **Products provide**
  - 40+ types of RS inversion products: vegetation structure and growth state, radiation budget, water and heat fluxes, ice-snow variation, mineral exploration
  - 20+ types of thematic RS products: agriculture, forestry, geology, water, sea ice, environment
  - large scales area: global, China, South America
  - long time series: all year round

- **Service provide**
  - spatial metadata management for multi-level RS data;
  - data-dependent knowledge base;
  - replicas management for distributed RS data;
  - automated data production;
  - production task scheduling;
  - workflow fault-tolerant;
  - distributed resource monitoring

RS file management

Lizhe Wang  HPC, Clouds, RS Big Data
RS Metadata management

Metadata Query

Bloom Filter Indexing

H(x1)  H(x2)  H(x3)  ...  H(xk)

BFV1  BFV2  BFVm

R-tree Indexing inside Groups

HBase Database
(RS Metadata)

Group 1 of entry files
Group 2 of entry files
Group m of entry files

HDFS

Storage

Data Center

Storage

Data Center

Storage

Data Center

Lizhe Wang  HPC, Clouds, RS Big Data
The Management System of Distributed RS Data Cache
1 RS Big Data: an Overview

2 How Big? a Quantitative Model

3 Software Systems for High Performance RS Big Data Processing
   - Global Distributed RS Data Model in Memory
   - Generic Parallel Processing Model for RS Big Data
   - Parallel Storage Model for RS Big Data
   - Scheduling of Large-scale RS images Parallel Processing

4 RS Big Data Processing in Clouds: Current Work

5 Concluding Remark
Concluding Remark for RS Big Data

Findings
- Big Data
- Compute-intensive and data-intensive
- Unbalanced Data Flow
- Complex Algorithm Flow

Recommendation
- Specially optimized HPC software systems for data transfer, IO, storage, processing and management
Thank you for your time!