EXTASCID: An Extensible System for the Analysis of Scientific Data
Yu Cheng, Florin Rusu
Electrical Engineering and Computer Science, University of California, Merced
ycheng4@ucmerced.edu, frusu@ucmerced.edu

System Architecture
EXTASCID is a parallel system targeted at efficient analysis of large-scale scientific data
- Massive heterogeneous data: data partitioning; parallel execution; relational and array data model
- Extensible complex analytics: user code executed inside the engine; enhanced UDA interface
- Architecture independence: multi-thread (shared memory and shared disk) and inter-node (shared nothing) parallelism

Storage Manager
Push-based execution: data streaming
Horizontal & vertical partitioning: column-oriented chunks
Dimension (index) suppression for dense grids

Processing Examples
Average computation
Clustering

Execution Engine
Merge-oriented parallel processing
Extended User-Defined Aggregate (UDA) interface

SS-DB Benchmark Experiments
System: 9-node cluster; 1 coordinator; 8 workers
Data loading and derived data computation

<table>
<thead>
<tr>
<th>System</th>
<th>Stage</th>
<th>Execution time [seconds]</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXTASCID Array</td>
<td>Compile</td>
<td>-240.5</td>
</tr>
<tr>
<td></td>
<td>Execute</td>
<td>1,823 82.2</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1,823 322.6</td>
</tr>
<tr>
<td>EXTASCID Relational</td>
<td>Compile</td>
<td>45.2</td>
</tr>
<tr>
<td></td>
<td>Execute</td>
<td>1,829 797.6</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1,829 842.6</td>
</tr>
<tr>
<td>SciDB</td>
<td></td>
<td>30,381 1,086</td>
</tr>
</tbody>
</table>

Queries on raw data

Queries on observations

Queries on groups of observations

Yu Cheng, Florin Rusu (EECS, University of California, Merced)
Extremely Large Databases (XLDB 2012)