Efficient Data Compression by Efficient Use of HDF5 Format with Bit Segmentation Method for Visualization and Analysis

Katsumi Hagita (National Defense Academy of JAPAN), Manabu Omiya (Hokkaido Univ.), Takashi Honda (Zeon), Masao Ogino (Nagoya Univ.)

Minimize file size for VR visualization.

Data Transfer Problem

Relatively narrow LAN and small disk space at local environment

Peta scale Storage

Concept of JHPCN-DF

Jointed Hierarchical Precision Compression Number - Data Format (JHPCN-DF)

Example of implementation

```
union fi32{
  float f;
  int i32;
};
union fi32 fival;
double logallo=log(allowerr)/log(2.0);
fival=frexp(fval0,&ival);
ival2=(int)(-logallo+ival-1); sval=(int)(23-ival2+1);
if(sval>23) sval=23;
do {
  sval--; fival.f=fval0;
  fival.i32=(fival.i32 >> sval);
  fival.i32=(fival.i32 << sval);
  fival1=fival.f;
} while (((fival1-fval0)*(fival1-fval0)>allowerr*allowerr));
```

Related method of standard HDF5

- N-bit Filter
  packing n-bit data on output by stripping off all unused bits and unpacking on input (restoring extra bits)
- Scale-Offset Filter
  Truncate value to a minimum number of bits before storing by scale and/or offset Relative error < 5*(D-scale_factor+1))

Advantage of JHPCN-DF

- Decodable without special API.
  • JHPCN-DF is effective for almost software and user application supporting HDF5.
- Splitting a few HDF5 files by simple converter application.
- Recoverable with lossless.

Reduction of Data Size

Data Size of zero padding is almost zero.

Huffman coding

This HDF5 file can be used by the standard API.
Acknowledgement

This work is partially supported by "Joint Usage/Research Center for Interdisciplinary Large-scale Information Infrastructures" and "High performance Computing Infrastructure" in Japan. 
JHPCN 14-NA28, HPCI hp140191 (JHPCN 10-MD01, 11-MD02, 12-MD03, jh130028-NA19, HPCI hp130056, hp130062, hp130122)

This work is partially supported by HPC project of Nagoya University.

Test for plasma PIC simulation

150,000 ions and 150,000 electrons

Comparison for allowerr =0.01, 0.02, 0.05

<table>
<thead>
<tr>
<th>exponent fraction</th>
<th>sign (8bit) (23bit)</th>
<th>exponent fraction</th>
<th>sign (8bit) (23bit)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>6,336 MB</td>
<td>0.001</td>
<td>3,369 MB</td>
</tr>
<tr>
<td>0.01</td>
<td>1,940 MB</td>
<td>0.02</td>
<td>1,413 MB</td>
</tr>
<tr>
<td>0.02</td>
<td>883.1 MB</td>
<td>0.05</td>
<td>568.5 MB</td>
</tr>
<tr>
<td>0.1</td>
<td>568.5 MB</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Test for FEM simulation

Finite element analysis of 0.24 million elements and 0.38 million nodes

<table>
<thead>
<tr>
<th>Size</th>
<th>150,000 ions and 150,000 electrons</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.3MB</td>
<td>Good for analysis</td>
</tr>
<tr>
<td>3.3MB for 1.0e-6</td>
<td>Good for visualization</td>
</tr>
<tr>
<td>1.3MB for 1.0e-3</td>
<td></td>
</tr>
</tbody>
</table>

Test for block copolymer

Reduce data size from 1182MB to 42MB to present isosurface of phase separated block copolymers. (mesh size: 512^3)

Size of 20 HDF5 files of 100 timesteps

Conclusion and Future

- Concept of JHPCN-DF works and shows excellent results for various type of simulation! JHPCN-DF is a powerful tool for all large scale simulations.
- Distribute know-how of JHPCN-DF for all supercomputing.
  - Increase number of examples for various simulation fields (in progress)
    - Nonhydrostatic ICosahedral Atmospheric Model (NICAM)
    - Fluid dynamics simulation
    - 0.1 billion powder simulation
- We will upgrade as a filter of HDF5.
- We consider post process of HDF5 using MPI-IO to be compressed by JHPCN-DF as system software coupled with job scheduler.
- Treatment of the last bit is related to rounding mode control of IEEE 754.