Distributing Higher-Dimensional Simulations Across Compute Systems: A Widely Distributed Combination Technique

Theresa Pollinger, Marcel Hurler, Michael Obersteiner, and Dirk Pflüger
Overview

1. Vision: Simulating the Numerical ITER
2. Advection with the Combination Technique
3. Hierarchical Communication Set-Up with DisCoTec
4. Numerical and Timing Measurements
5. Conclusion and Future Work
Vision: Simulating the Numerical ITER

ITER and Numerical ITER with GEN\(E\), Source: [1], [2]
Proof-of-Concept Scenario: 6D-Advection

Incompressible advection equation

\[ \frac{\partial u}{\partial t} + \vec{v} \cdot \nabla u = 0 \]  \hspace{1cm} (1)

on \( \Omega = [0, 1]^d \), \( d \) dimensions, periodic boundary conditions in space

\[ u(\vec{x}, t = 0) = \exp\left(- \sum_{i=1}^{d} (x_i - \frac{1}{2})^2 \cdot \frac{1}{\sigma^2}\right), \sigma = \frac{1}{3} \]  \hspace{1cm} (2)

solve by backward finite differences in both space and time

\[ \Rightarrow \text{explicit Euler time-stepping scheme:} \]

\[ u(\vec{x}, t) = u(\vec{x}, t - \Delta t) - \nabla u \cdot \vec{v} \cdot \Delta t. \]  \hspace{1cm} (3)
The Sparse Grid Combination Technique

Truncated sparse grid combination technique with $\vec{\ell}_{\min} = [2, 1]$ and $\vec{\ell}_{\max} = [5, 4]$
DisCoTec-Simulations

DisCoTec: Run-combine cycle on three process groups.
DisCoTec-Simulations

DisCoTec-Simulations

Communication Between HPC Systems

Network

HPC System 1  HPC System 2

Direct communication between HPC systems impossible!
Communication Between HPC Systems

Direct communication between HPC systems impossible!
Widely-Distributed Combination Technique

Distributing Higher-Dimensional Simulations Across Compute Systems: A Widely Distributed Combination Technique

Theresa Pollinger
Numerical Errors

Monte-Carlo approximation on $1 \times 10^5$ samples

$$\| u^c - u^* \|_2(t) \approx \sum_{j=1}^{N} \frac{(u^c(x_j, t) - u^*(x_j, t))^2}{N},$$

$$\| u^c - u^* \|_0(t) \approx \max_{x_j} (u^c(x_j, t) - u^*(x_j, t))$$
**Numerical Errors**

Monte-Carlo approximation on $1 \times 10^5$ samples

$$\| u^c - u^* \|_2(t) \approx \frac{1}{N} \sum_{j=1}^{N} \frac{(u^c(x_j, t) - u^*(x_j, t))^2}{N},$$

$$\| u^c - u^* \|_0(t) \approx \max_{x_j} (u^c(x_j, t) - u^*(x_j, t))$$

Errors for advection problem over time. Results are the same for the locally and widely distributed cases.
Local Weak Scaling Set-Up
Local Weak Scaling Set-Up

⇒ shifting scenario + doubling worker processes: (almost) constant work
Local Weak Scaling Set-Up

⇒ shifting scenario + doubling worker processes: (almost) constant work

Scenario for 1 process group of size 1

- $\ell^\text{min} = [2 1 1 1 1 1]$, $\ell^\text{max} = [9 8 8 8 8 8] \Rightarrow 1709$ component grids
- $\approx 1.153 \text{ GiB memory for full grid data} \ (7.74 \times 10^7 \ \text{DOF} \cdot 2)$
- $\approx 19.25 \text{ MiB memory for sparse grid data}$
Weak Scaling Results

Each experiment is averaged over 10 run/combine cycles; bars = standard deviation.
## Strong Scaling Set-Up

### Scenario

same as before for 64 processes in the process group:

- $\ell^{\text{min}} = [3\ 2\ 2\ 2\ 2\ 2]$, $\ell^{\text{max}} = [10\ 9\ 9\ 9\ 9\ 9]$
- $>37$ GiB memory for full grid data ($2.53 \times 10^9$ DOF $\cdot 2$)
- $\approx 673.84$ MiB memory for sparse grid data
## Strong Scaling Set-Up

### Scenario

same as before for 64 processes in the process group:

- \( \ell_{\text{min}} = [3 2 2 2 2 2], \ell_{\text{max}} = [10 9 9 9 9 9] \)
- \( > 37 \text{ GiB memory for full grid data (2.53 } \times 10^9 \text{ DOF } \cdot 2 \)
- \( \approx 673.84 \text{ MiB memory for sparse grid data} \)

\[ \Rightarrow \text{ keep problem size constant, add process groups [4], [5]} \]
Local Strong Scaling Results

Each experiment averaged over 10 run/combine cycles and all process groups, bars = standard deviation. Load imbalance sets in as 1709 grids are distributed among 1023 process groups (i.e. 65472 processes).
Remote Communication Set-Up

Network

Machine A → Broker → Machine B

Image courtesy [6]
Remote Communication Set-Up

<table>
<thead>
<tr>
<th>Network</th>
<th>Machine A</th>
<th>Broker</th>
<th>Machine B</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th></th>
<th>nodes</th>
<th>cores/node</th>
<th>main memory</th>
<th>processor model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine A (HAWK [7])</td>
<td>5632</td>
<td>128</td>
<td>2 GiB/core</td>
<td>AMD EPYC 7742 “Rome”</td>
</tr>
<tr>
<td>Machine B (local)</td>
<td>1</td>
<td>128</td>
<td>16 GiB/core</td>
<td>AMD EPYC 7742 “Rome”</td>
</tr>
</tbody>
</table>

Image courtesy [6]
Widely Distributed Strong Scaling Results

Overheads of the combination step for the remote strong scaling case. Each data point is averaged over 4 experiments, each experiment is averaged over 10 run/combine cycles. Error bars = standard deviation. Measurements are taken on machine A’s manager process.
The plasma solver lines are estimates based on a reference run – a single time step takes about four times as long as with the advection solver.
Conclusion and Future Work

• Benefits: Accelerate large-scale computations, run scenarios that would exceed one system’s memory
Conclusion and Future Work

- Benefits: Accelerate large-scale computations, run scenarios that would exceed one system’s memory
- In this case: all of the drawbacks, none of the benefits
- Will pay off for large scenarios w/ more complicated solvers
Conclusion and Future Work

- Benefits: Accelerate large-scale computations, run scenarios that would exceed one system’s memory
- In this case: all of the drawbacks, none of the benefits
- Will pay off for large scenarios w/ more complicated solvers
- Asynchronous combination [8] hides communication latency
Conclusion and Future Work

- Benefits: Accelerate large-scale computations, run scenarios that would exceed one system's memory.
- In this case: all of the drawbacks, none of the benefits.
- Will pay off for large scenarios w/ more complicated solvers.
- Future work: two large systems, more physics.
References I


M. Obersteiner, “A spatially adaptive and massively parallel implementation of the fault-tolerant combination technique”, Dissertation, Technische Universität München, To Be Published.