Memory-Aware Scheduling Of Tasks Sharing Data On Multiple GPUs
Maxime Gonthier, Samuel Thibault, Loris Marchal

To cite this version:
Maxime Gonthier, Samuel Thibault, Loris Marchal. Memory-Aware Scheduling Of Tasks Sharing Data On Multiple GPUs. HiPEAC ACACES 2022 - 18th International Summer School on Advanced Computer Architecture and Compilation for High-performance Embedded Systems, Jul 2022, Fiuggi, Italy. hal-04090607

HAL Id: hal-04090607
https://inria.hal.science/hal-04090607
Submitted on 5 May 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Motivation: Extract peak performance from GPUs

- Scientific application uses a huge quantity of data
- GPUs can calculate quickly but their memory are limited
- The PCI bus becomes a bottleneck when large amounts of data need to be loaded

Objective 1: Minimize the number of tasks on each GPU
Objective 2: Minimize the amount of data loads on each GPU

A novel strategy: Data-Aware Reactive Task Scheduling (DARTS)

Flow of tasks

<table>
<thead>
<tr>
<th>PlannedTasks</th>
<th>TaskBuffer</th>
<th>GPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
</tbody>
</table>

- Independent tasks sharing data
- Same size data
- Multiple GPUs with limited memory

Framework and objectives

- Independent tasks sharing data
- Same size data
- Multiple GPUs with limited memory

Figure: Example with a memory of size 2 data. The graph of input data dependencies is shown on the left. The figure on the right corresponds to the partition and schedule produced by the scheduler.

Algorithms from the state of the art

- Using (hyper-)graph partitioning: hMETIS
  - Hypergraph \( \rightarrow \) Represents a data being shared by multiple tasks
  - hMETIS produces subsets of task that maximize the amount of edges (data shares) inside each subset
  - We add dynamic reordering and task stealing to improve performance

- Deque Model Data Aware Ready (DMDAR):
  - Schedules tasks where their completion times are expected to be minimal. Uses a ready strategy to favor tasks whose data has already been loaded into memory.

- EAGER
  - Process tasks in the order in which it receives them \( \rightarrow \) our baseline

- DMDAR suffers from a conflict between prefetches and evictions

- DARTS+OPTI is a variant that reduces large scheduling cost for all algorithms

- hMETIS suffers from a large scheduling cost

Experimental settings

- Tesla V100 GPUs
- 12000 MB/s PCI bandwidth
- GPU memory limited to 500 MB to better distinguish the performance of different strategies even on small datasets

Table: CUMULATED GPUs MEMORY

<table>
<thead>
<tr>
<th>GPUs</th>
<th>Memory (GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>500</td>
</tr>
<tr>
<td>2</td>
<td>1000</td>
</tr>
<tr>
<td>3</td>
<td>1500</td>
</tr>
<tr>
<td>4</td>
<td>2000</td>
</tr>
</tbody>
</table>

Summary and conclusion

- Limiting data movement is crucial to extract the most out of GPUs
- We provide a new algorithm (DARTS) and a new eviction policy (LUF) focused on data locality

- Results: DARTS+LUF achieves good performance because it:
  - Limits data transfers thanks to the finding of an optimal data and an adapted eviction policy
  - Overlaps communication and computations by distributing transfers over time
  - Can be used with a reduced complexity

- Areas for improvement:
  - Consider tasks with dependencies
  - Consider inter-GPU communications (NVLink)
  - Manage multiple MPI nodes

Figure: Visualization of DARTS processing order on the tiled Matrix Multiplication

Figure: tiled sparse Matrix Multiplication in outer product order without memory limitation (32GB by GPU) with 4 Tesla V100 GPUs

Figure: Cholesky decomposition (without dependencies) with 4 Tesla V100 GPUs

Figure: PCIs (left) and data transfers (right) on a tiled \( (1 \times n) \) Matrix Multiplication in outer product order with 2 Tesla V100 GPUs.