The Locality Descriptor

A Holistic Cross-Layer Abstraction to Express Data Locality in GPUs

ISCA 2018

Nandita Vijaykumar

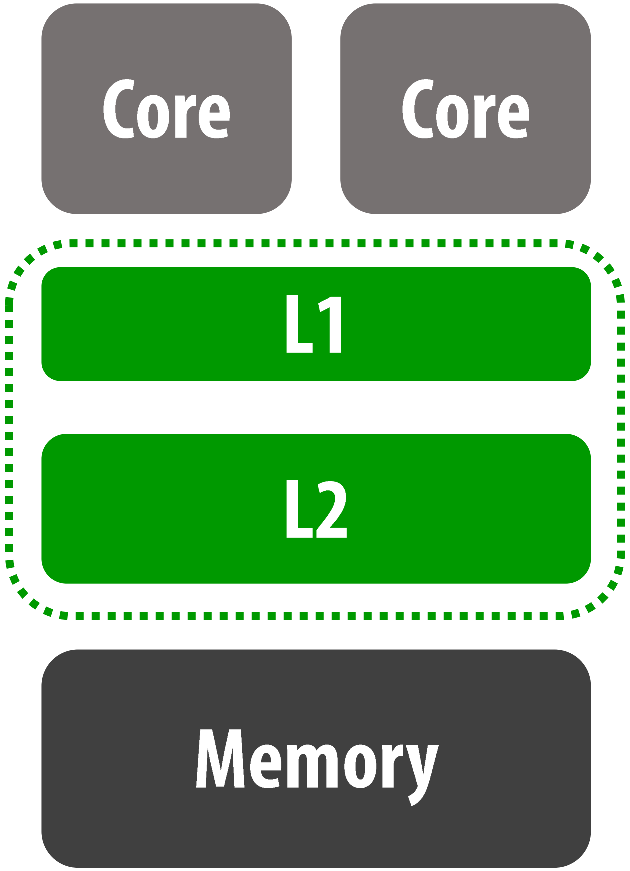
Eiman Ebrahimi, Kevin Hsieh, Phillip B. Gibbons, Onur Mutlu

Carnegie Mellon University

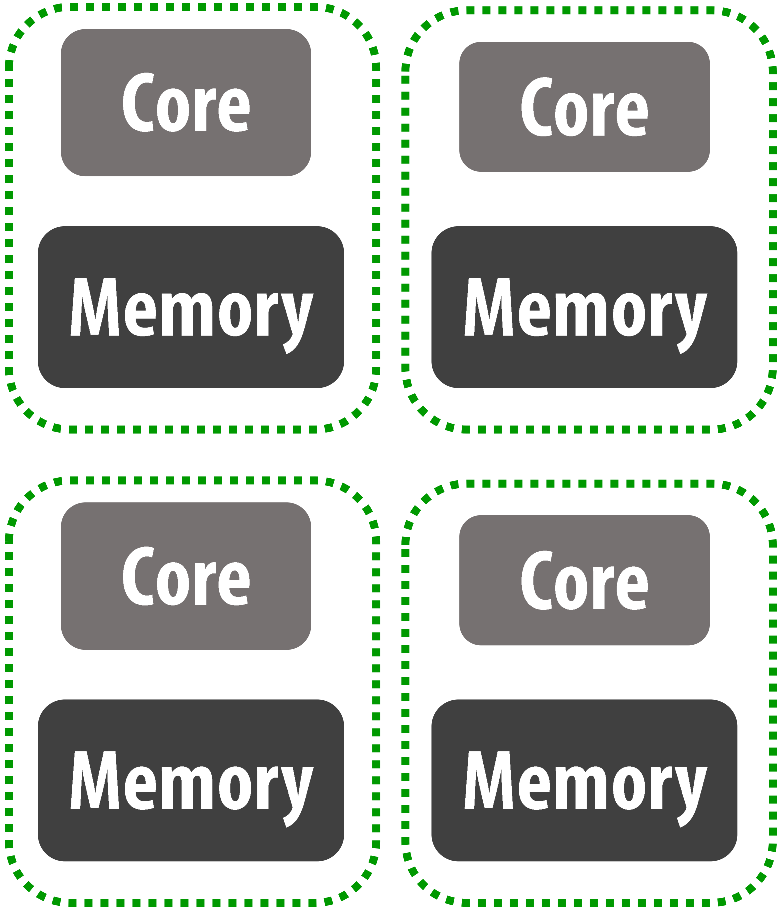
NVIDIA

ETH Zürich

Data locality is critical to GPU performance



Cache Locality



NUMA Locality

Exploiting data locality in GPUs is a challenging and elusive feat…

…requiring a range of architectural techniques

Cache Management

CTA Scheduling

Data Placement

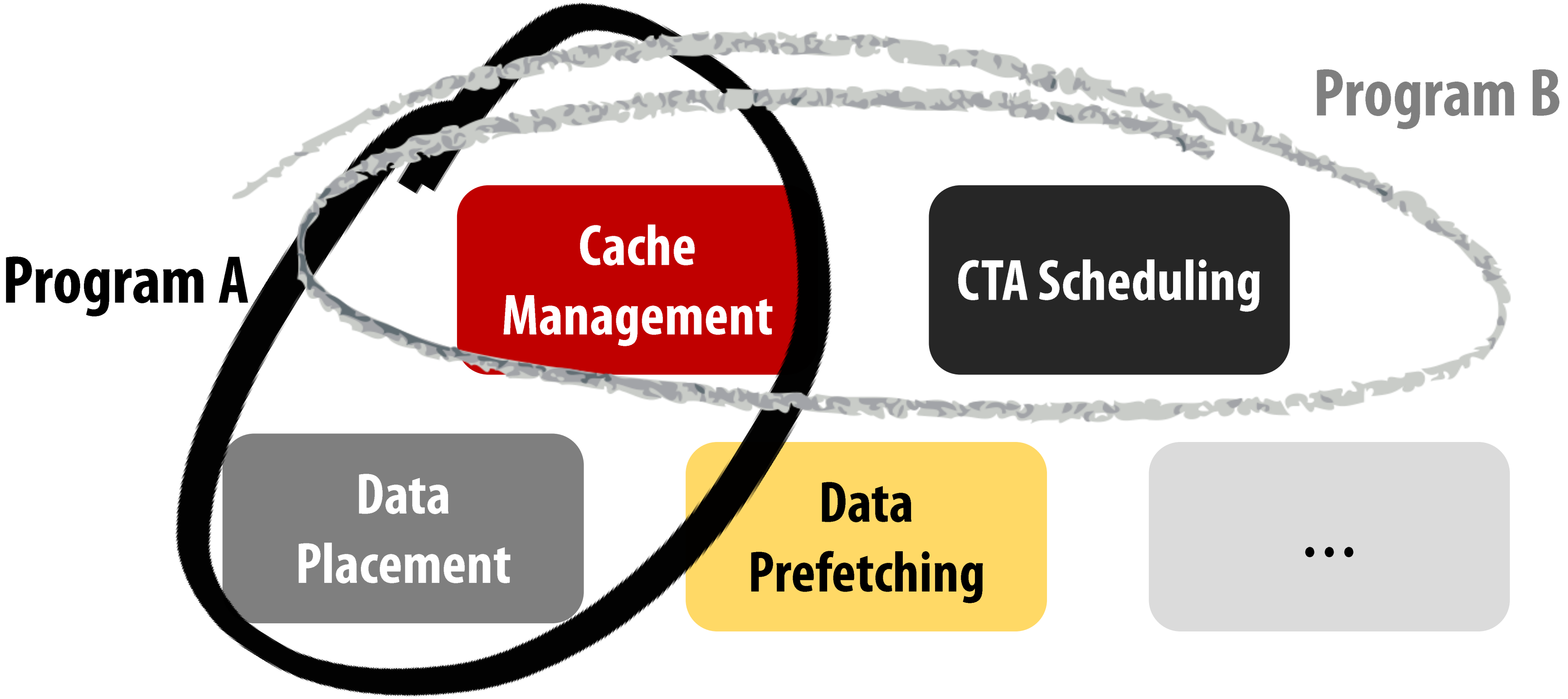
Data Prefetching

…

Furthermore…

A single technique is often insufficient

The required set of techniques depends on the program



Challenging for the programmer/software

No easy access to many architectural techniques

Tedious and un-portable programming:

```
Bypass Cache Line A
Schedule Thread Block 2 at SM 1
...
```

Challenging for the architect

Hardware misses key program semantics required for optimization

Where to place data?

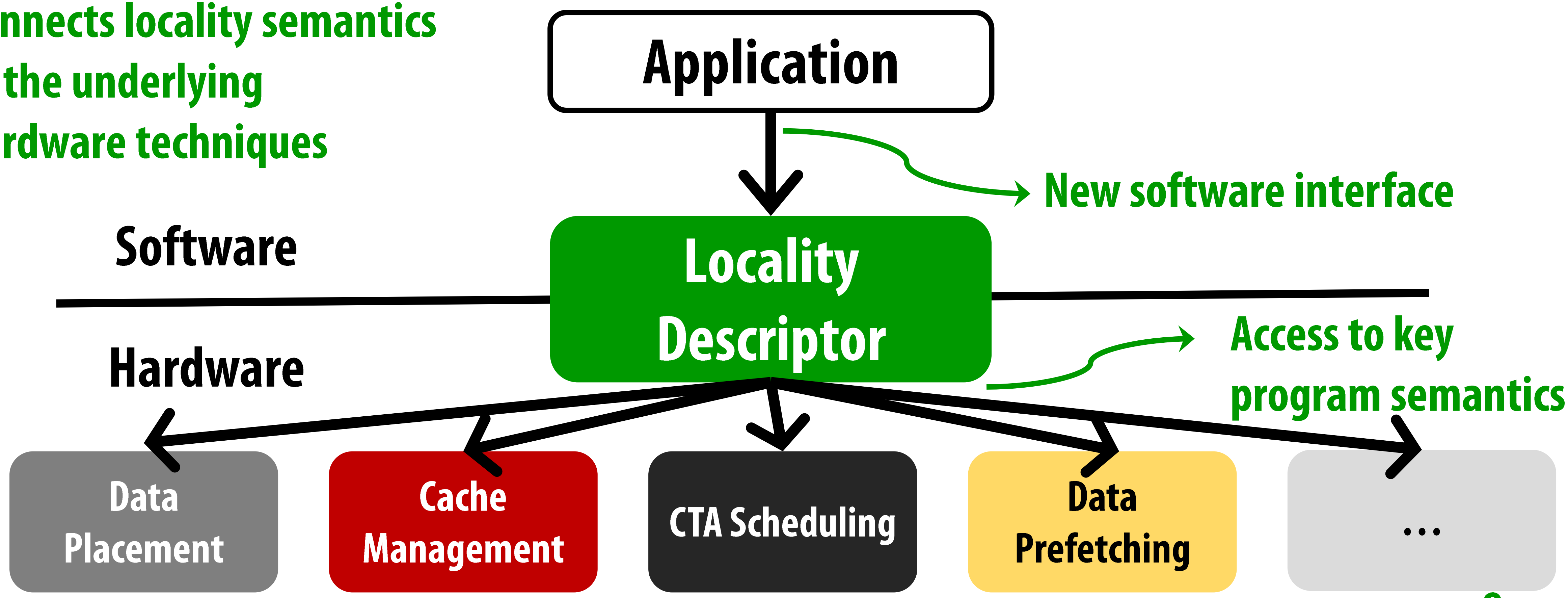
Which threads to schedule together?

Which data to bypass?

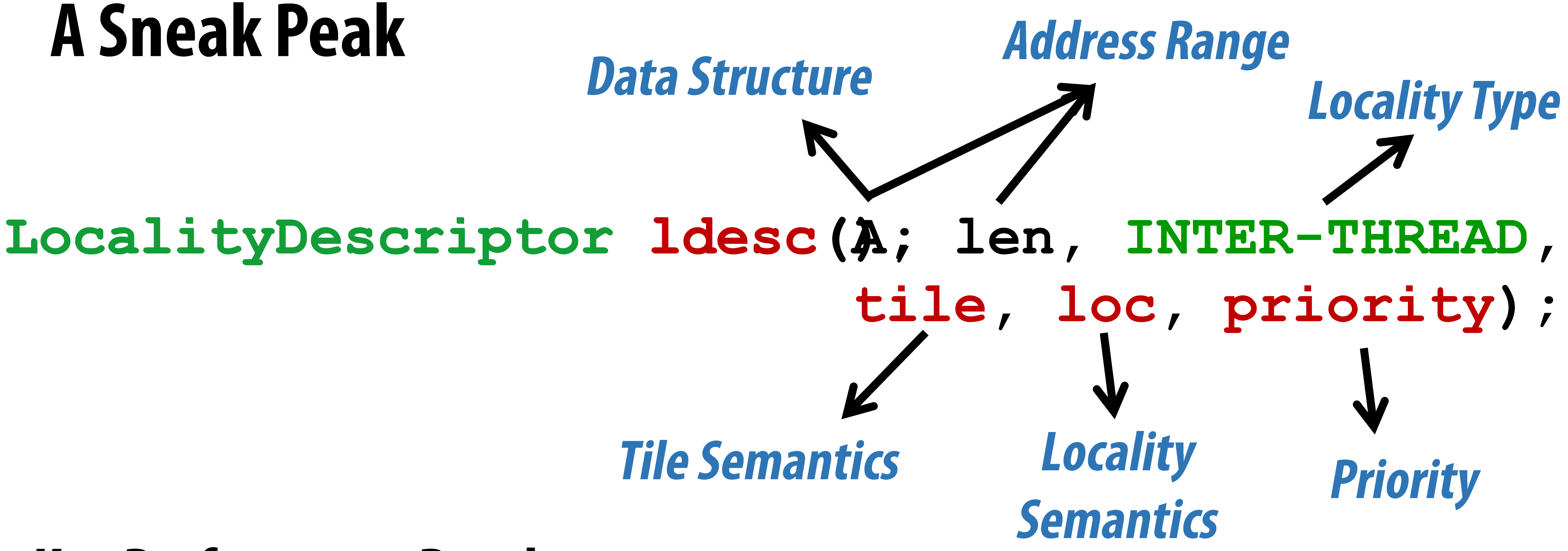
The Locality Descriptor

A hardware-software abstraction to express and exploit data locality

Connects locality semantics to the underlying hardware techniques



A Sneak Peak



```
LocalityDescriptor ldesc(A; len, INTER-THREAD,
                         tile, loc, priority);
```

Data Structure

Address Range

Locality Type

Tile Semantics

Locality Semantics

Priority

Key Performance Results:

Leveraging Cache Locality: ↑26.6% on average (up to 46.6%)

Leveraging NUMA Locality: ↑53.7% (up to 2.8X)

The Locality Descriptor

A Holistic Cross-Layer Abstraction to Express Data Locality in GPUs

ISCA 2018

Nandita Vijaykumar

Eiman Ebrahimi, Kevin Hsieh, Phillip B. Gibbons, Onur Mutlu

Carnegie Mellon University

NVIDIA

ETH Zürich