

# Sorting using Bitonic Network with CUDA

Gabriele Capannini, Franco M. Nardini,  
Fabrizio Silvestri and Ranieri Baraglia

{gabriele.capannini, francomaria.nardini,  
fabrizio.silvestri, ranieri.baraglia}@isti.cnr.it

Information Science and Technologies Institute (ISTI) of the  
Italian National Council (CNR)



# Outline

- ISTI's HPCLab
- Compute Unified Device Architecture (CUDA)
- Bitonic sort with CUDA

# HPC Lab @ ISTI

- 8 Researchers
- 2 Technicians
- 1 Administrative
- 3 PhD fellows
- 1 MsD fellow
- 4 PhD students
- 5 Research collaborators



# Research activities @ ISTI-HPCLab

- Research on systems for computational and data-intensive problems...
- ... in business, social, scientific, and knowledge-based applications...
- ... dealing with the exponential growth in the amount of services, data, knowledge and users

**HPC solutions for efficient and scalable management, summarization, and search of data/services**

# Research topics @ ISTI-HPCLab

- **Scalable Data and Web Mining**
  - Scalable Pattern Mining
  - Query Log Analysis for efficient IR
  - Web recommender systems
- **Efficiency in Web Search**
  - Caching, Index partitioning and query routing
- **Many-core architectures**
  - Many-core algorithms for WIR and chemical applications
- **P2P Systems**
  - Resource/services discovery, search audio-visual content
- **Grid Computing**
  - Job scheduling, Mapping of parallel application, Grid operating systems
- **Parallel/Distributed Programming Environments**
  - Design of component based frameworks for grid applications.

## European Projects

Type	Name	Goal	HPCLab Role	Period
<b>FP6 NoE</b>	<b>S-Cube</b>	To push the frontiers of research in Service Oriented Computing	<i>Grid computing techniques for scalable and self-* service based applications infrastructures</i>	<b>2008-2013</b>
<b>FP6 STREP</b>	<b>Sapir</b>	Design P2P architecture for search audio-visual content using the query-by-example paradigm.	<i>Push-based crawling, scalable P2P indexing and caching for multimedia content</i>	<b>2007-2009</b>
<b>FP6 IP</b>	<b>XtreemOs</b>	Building of an open source Grid operating system	<i>Resource discovery from a P2P perspective</i>	<b>2006-2010</b>
<b>COST Action</b>	<b>HPC on Complex Environments</b>	Building of models, algorithms, programming tools and applications for novel hierarchical and heterogeneous computing platforms.	<i>Building of algorithms/applications for WIR, Computational chemistry (WG Applications)</i>	<b>2009-2013</b>
<b>FP7 IP</b>	<b>ASSETS</b>	To improve the usability of European digital library by building large-scale services focusing on search, browsing and interfaces	<i>Effective metadata ranking methods and design of large-scale search services</i>	<b>2009-2012</b>

## National Projects

Type	Name	Goal	HPClab Role	Period
<b>CNR-RSTL</b>	<b>Resource Discovery in P2P Networks</b>	To build new resource discovery tools for large P2P collaborative environment	<i>Study P2P models and protocols to support the realization of novel scalable discovery tools, with different mechanisms for different levels of dynamicity</i>	<b>2009-2010</b>
<b>Regional</b>	<b>VISITO</b>	To design fully immersive virtual services supporting tourists before, during, and after their visits in Tuscany	<i>Trusted metadata harvesting, scalable multimedia management and retrieval, touristic recommendation algorithms</i>	<b>2009-2011</b>
<b>Industry</b>	<b>MOTUS</b>	To devise innovative services for mobility and tourism in urban scenarios.	<i>Scalable data management and recommendation algorithms for tourists</i>	<b>2009-2011</b>

## Main results (2006 – 2008)

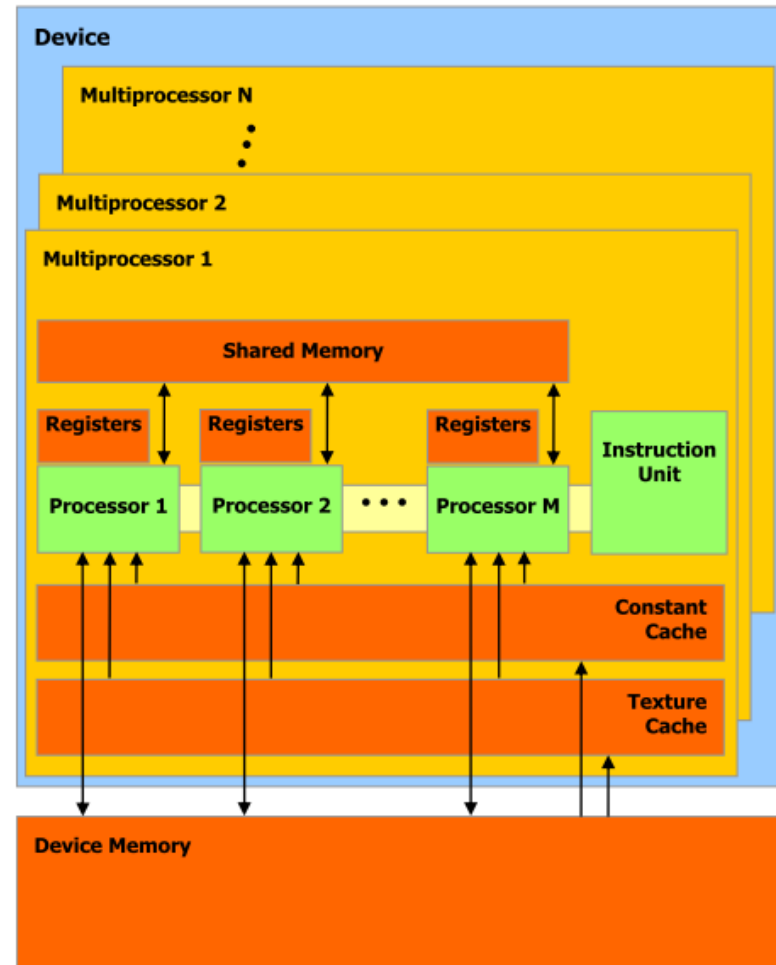
- International Journals 16
- International Conferences 66
- Book chapters 8
- Patents 3
- International conferences organization 92
- PhD students 10
- EC FP6/FP7 projects. 10
- National projects 5
- Research contract 2
- External funds € 1,3M

# Compute Unified Device Architecture

- CUDA is a hardware and software architecture for issuing and managing computations on the GPU.
- The CUDA programming model
  - Minimal extension of C and C++ languages
  - The programmer writes a serial program that calls parallel *kernels*
  - Serial portions execute on the host CPU
  - A kernel executes as parallel threads on the GPU
    - Kernel may be simple functions or full programs
    - Many threads execute each kernel
- Available for the GeForce 8 Series, the Tesla solutions, and some Quadro solutions.

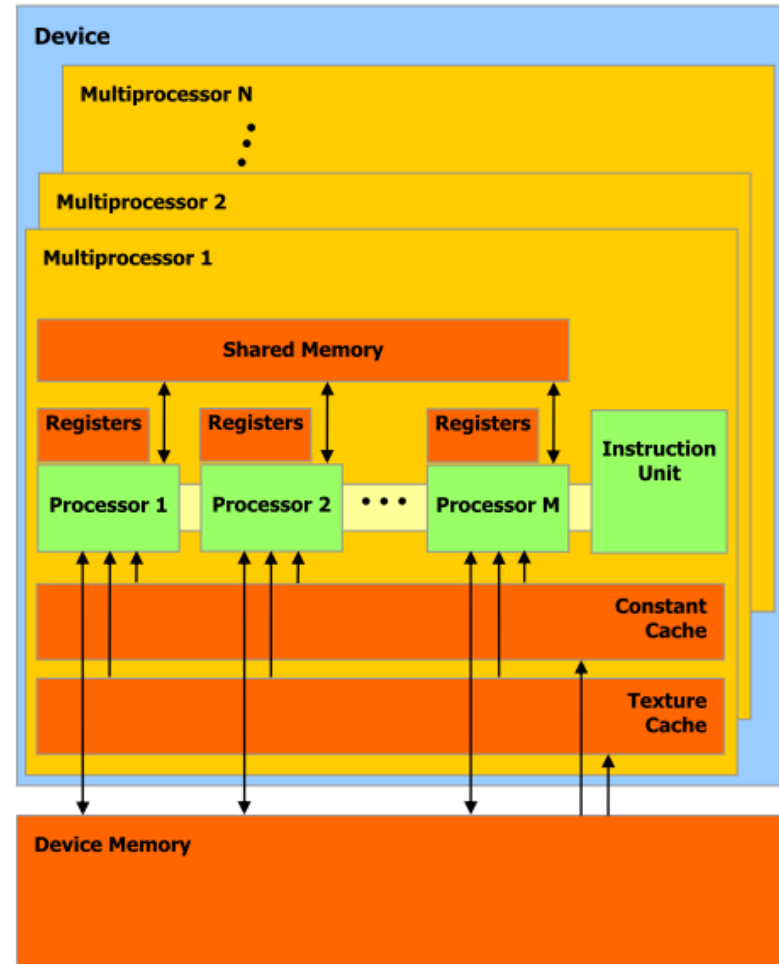
# CUDA Architecture

- **Thread**
  - Compute results elements
  - Identified by an id number (*threadIdx*)
  - For fine-grain data parallelism
- **Thread Block**
  - An array of threads that can cooperate by
    - sharing data through shared memory
    - synchronizing their execution
  - Compute result data block
  - Identified by an id number (*blockIdx*)
  - For coarse-grained data parallelism
- **Grid of Blocks**
  - Compute many result blocks
  - 1 to many blocks for grid
  - Identified by grid and block dimensions
  - For task parallelism
- **Sequential grids**
  - Compute sequential problem steps



# CUDA Architecture

- **Local memory per thread**
  - Private per thread
- **Shared Memory per Block**
  - Shared by threads of a block
  - Inter-threads communication
  - Barrier synchronization
- **Global memory per application**
  - Shared by all threads
  - Inter-grid communication
  - Inter-kernel synchronization



# Bitonic Sorting Network on GPU