“The Beach Law” does not hold any more

Discrete optimization needs heterogeneous computing

Christian Schulz, Trond Hagen, Geir Hasle
Department of Applied Mathematics, SINTEF ICT, Oslo, Norway

Seminar

CORAL, Aarhus School of Business and Social Sciences
Aarhus, Denmark, March 16, 2011

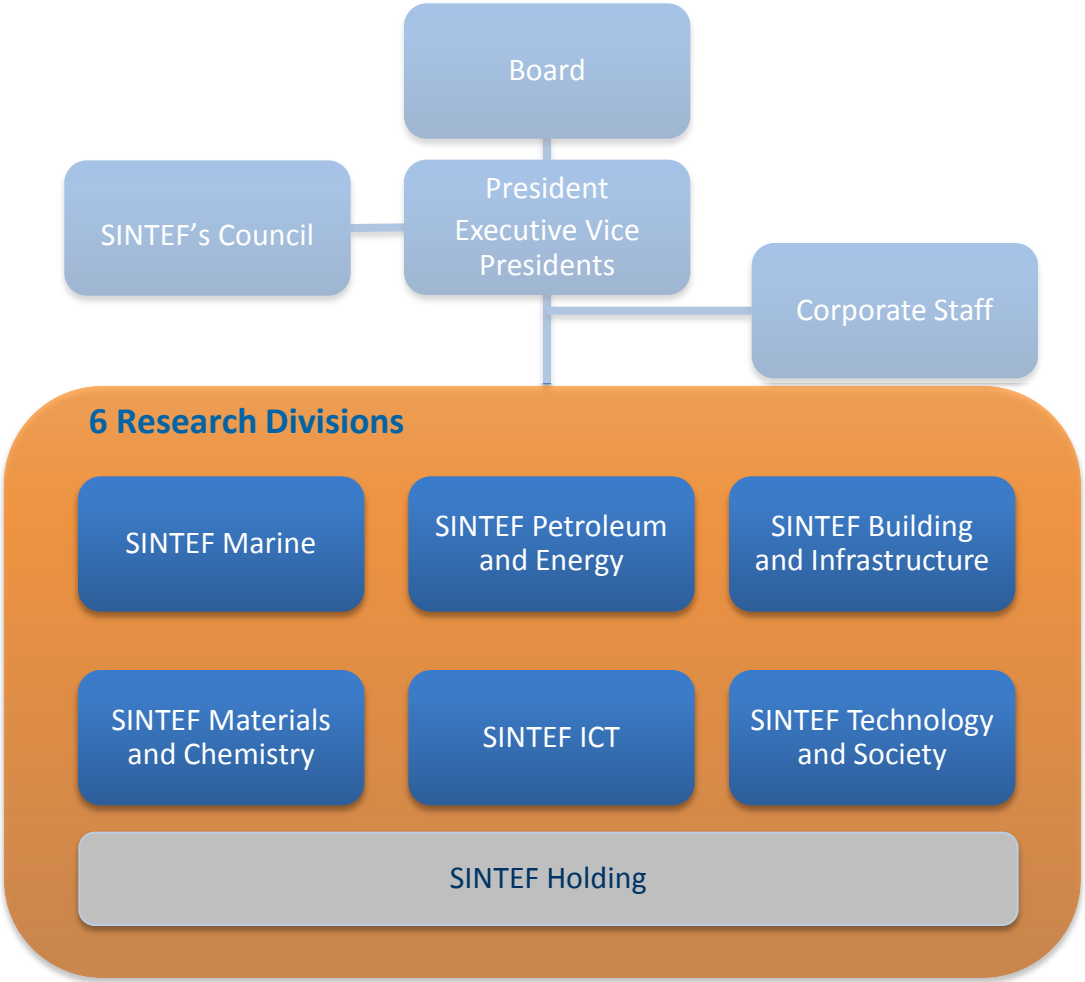
Outline

- SINTEF
- Performance in Discrete Optimization
- Hardware developments, and prospects
- Accelerators and heterogeneous computing
- «Camel Spider (Solifugae)» - a GPU based VRP solver
- Extension to truly heterogeneous computing
- Conclusions

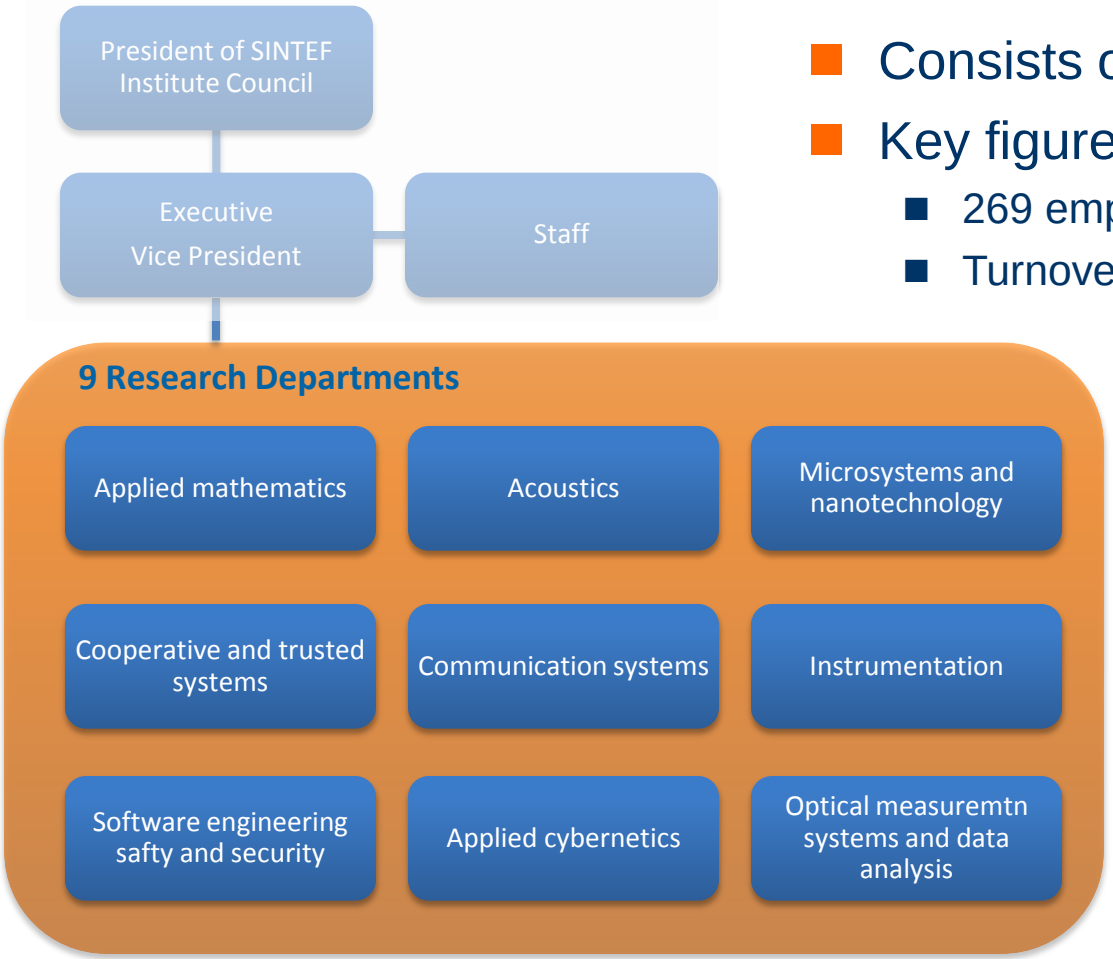
SINTEF

- Established 1950 by the Norwegian Institute of Technology.
- The largest independent research organisation in Scandinavia.
- A non-profit organization.
- Vision: "Technology for a better society".
- Key figures*
  - 2123 employees from 67 different countries.
  - 2755 MNOK turnover (about € 340M).
  - 7216 projects for 2200 customers.
  - Main offices in Trondheim and Oslo
  - Offices in USA, Brazil, Macedonia, United Arab Emirates, and Denmark.

The SINTEF Group

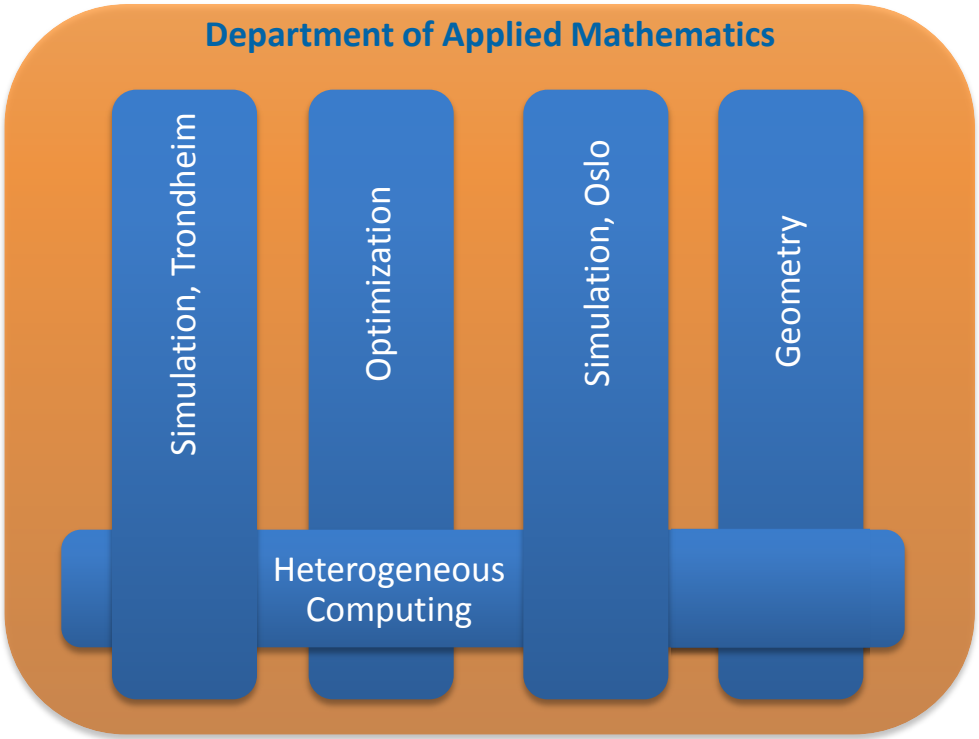


SINTEF ICT: Organization



- Offices in Oslo and Trondheim
- Consists of 9 departments
- Key figures 2009
  - 269 employees
  - Turnover 336 million NOK

Department of Applied Mathematics



- Offices in Oslo and Trondheim
- Consists of 5 research groups
  - Geometry
  - Optimization
  - Simulation
  - Visualization
  - Heterogeneous computing
- Key figures 2009
  - 38 employees
  - 45 MNOK turnover

SINTEF

Technology for a better society

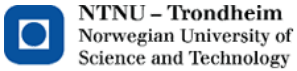
Optimization group

Overview: Optimization group

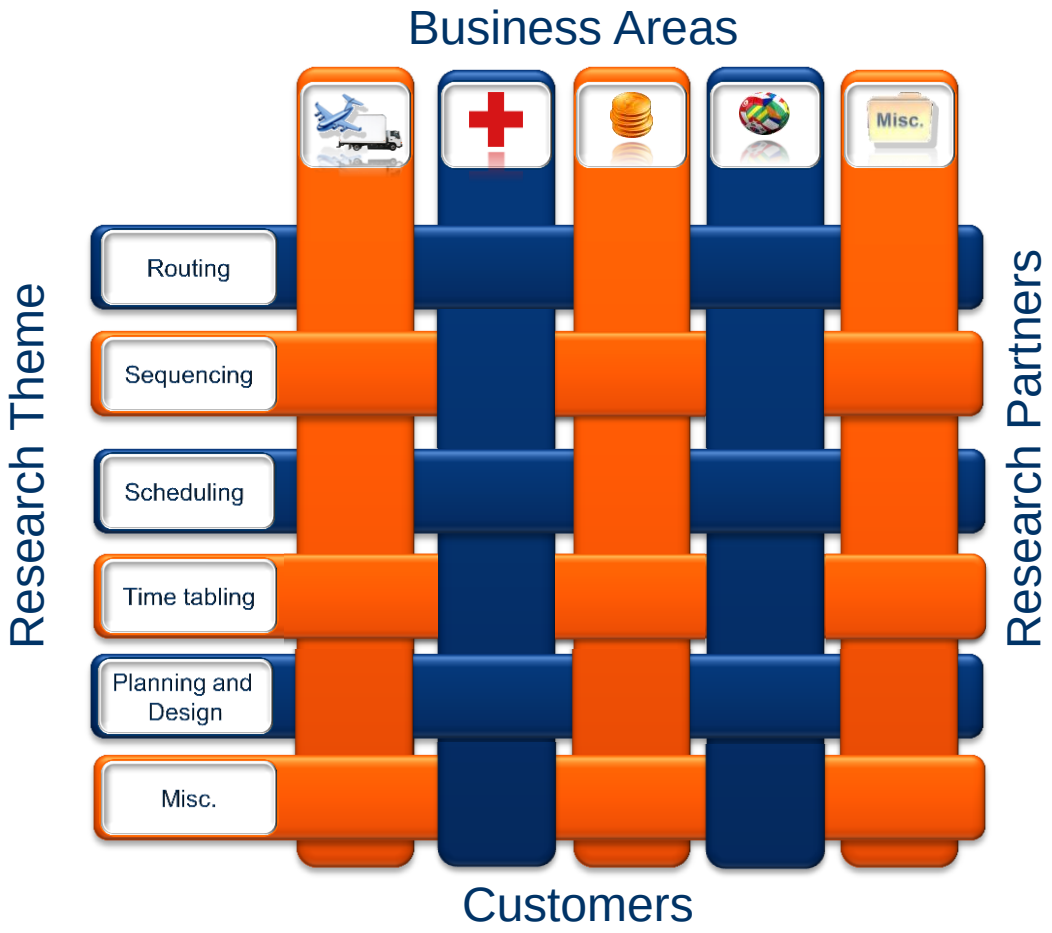
- **Focus**
  - 20 years of applied research in discrete optimization
- **Employees**
  - 9 researchers, 1 software engineer
- **Activities**
  - Basic research
  - Applied research
  - Consultancy
- **Products and Services**
  - Develop mathematical optimization models and algorithms
  - Develop optimization software (stand alone and plugin)
    - SCOOP library
    - Spider VRP Solver, Invent inventory routing solver
  - Consultancy reports, scientific papers
  - Spin-outs

Customers and Partners

- **Customers**
  - Industry
  - Public Sector
  - Research Council of Norway
  - Norwegian Research bodies
  - European Union
- **Research Partners**
  - Industry
  - University
  - Research Institute



Business Areas & Research Themes



Outline

- SINTEF
- Performance in Discrete Optimization
- Hardware developments, and prospects
- Accelerators and heterogeneous computing
- «Camel Spider (Solifugae)» - a GPU based VRP solver
- Extension to truly heterogeneous computing
- Conclusions

Performance in Discrete Optimization

- DOPs computationally hard
- Tremendous increase in DOP solving ability
- Illustration: Commercial LP solvers*
- Speedup factor roughly 1.000.000 1987-2000
- Factor 1000 better methods
- Factor 1000 faster computers

- There is still a performance bottleneck in industry

*Bixby R.E. (2002). Solving Real-World Linear Programs: A Decade and More of Progress. Oper. Res. 50(1), pp. 3-15.

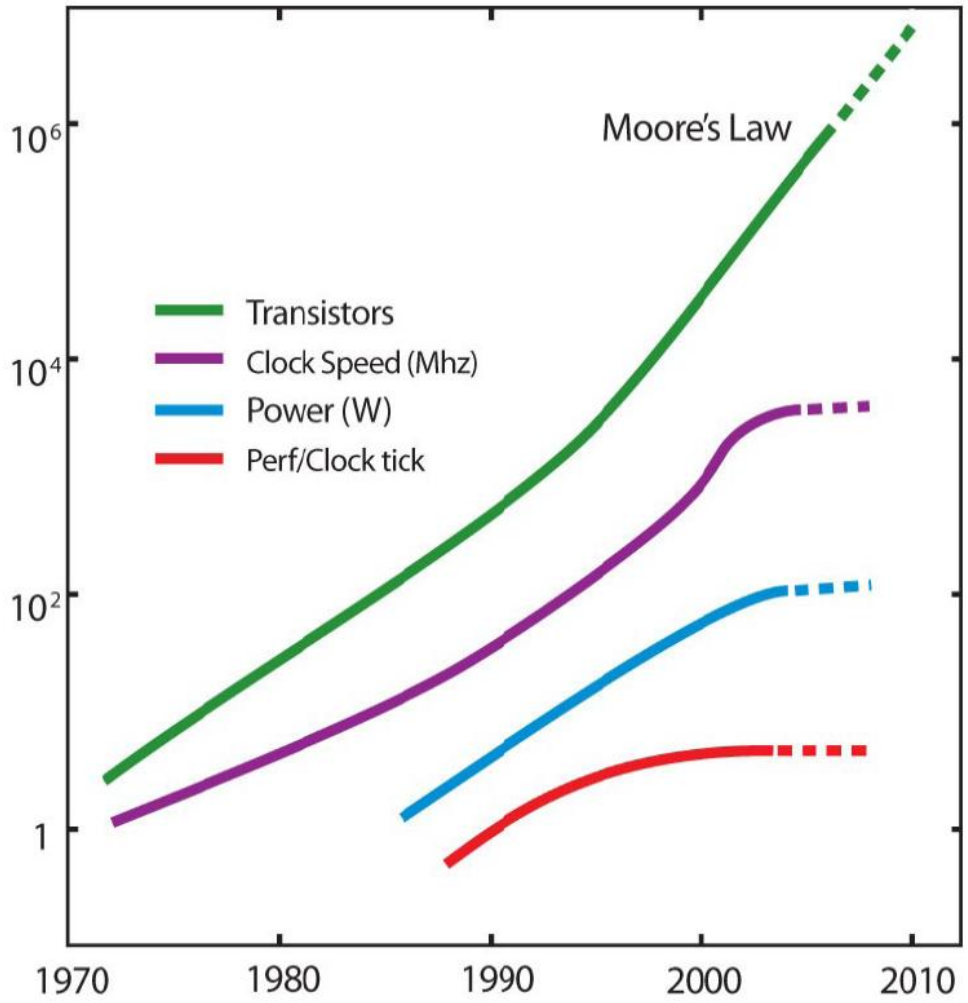
Outline

- SINTEF
- Performance in Discrete Optimization
- Hardware developments, and prospects
- Accelerators and heterogeneous computing
- «Camel Spider (Solifugae)» - a GPU based VRP solver
- Extension to truly heterogeneous computing
- Conclusions

The Beach Law [Gottbrath et al. 1999]

One way of doubling the performance of your computer program is to go to the beach for 2 years and then buy a new computer.

Processor development 1970-2010



"The number of transistors on an integrated circuit for minimum component cost doubles every 24 months"

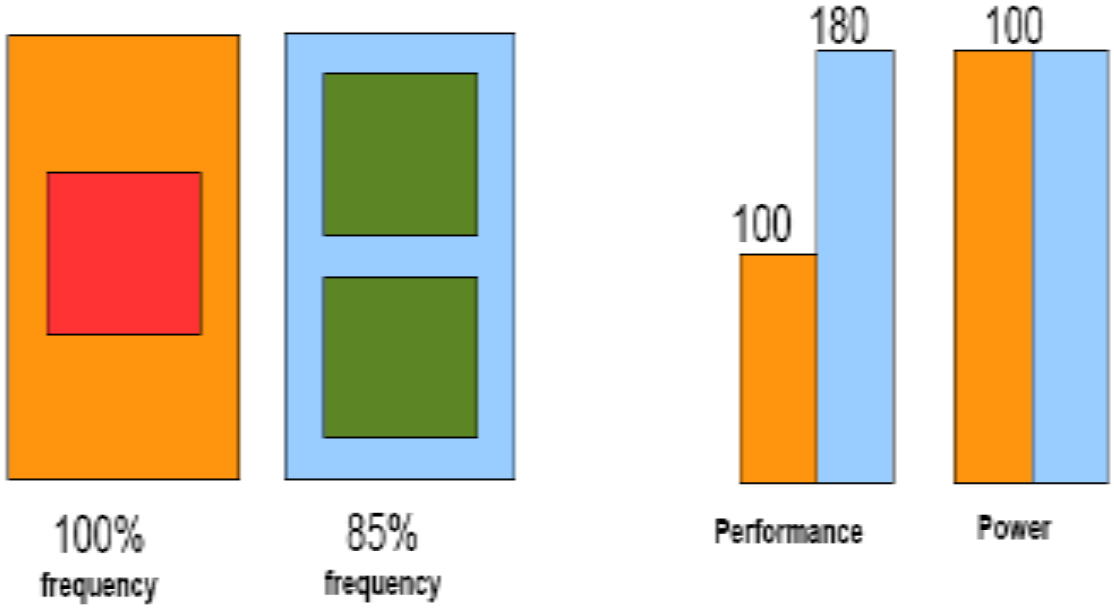
– Gordon Moore, 1965.

What happened?

- Moore's law at work, expected to hold until 2030 ...
- The Beach Law was valid until about 2005 ...
- Heat dissipation etc. stopped it
- PC computing power still benefits from Moore's law
- Multi-core processors for task parallelization (multi-threading, shared memory)
- Accelerators for data parallelization (stream processing)

- Drastic change in the development of processors

Multi-core processors



- Heat dissipation varies with clock frequency cubed
- 2 cores, reduced frequency, same heat dissipation
- 70% higher computing performance **if you can exploit it**
- **Sequential programs will run slower**

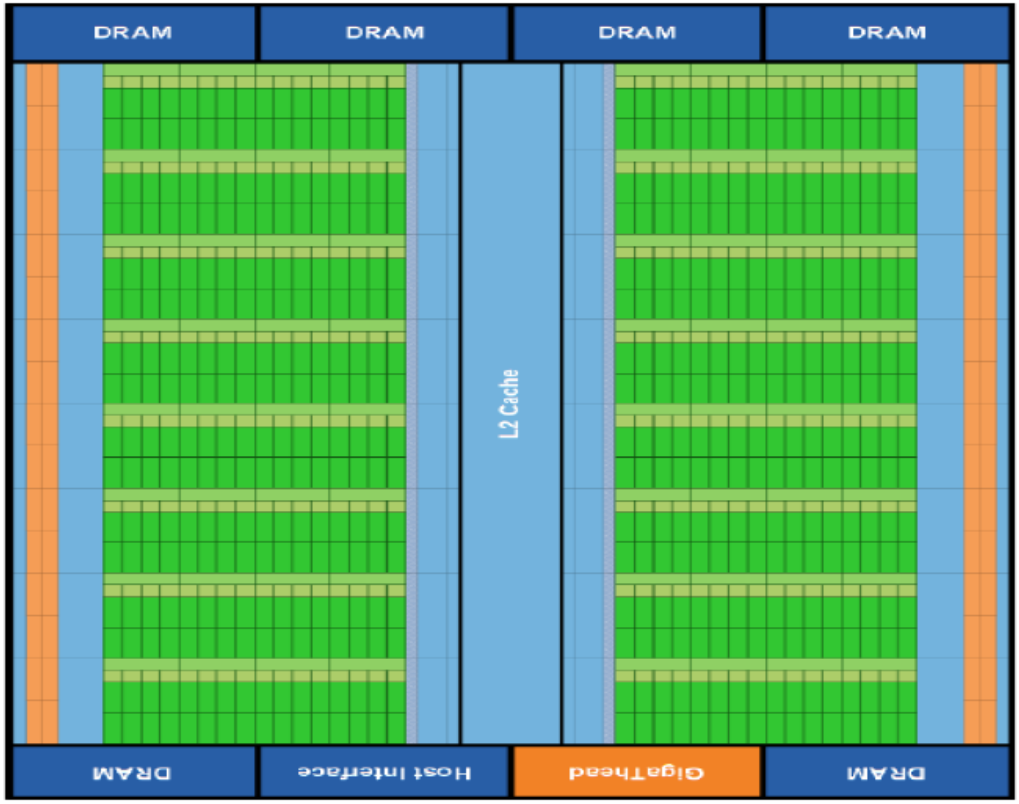
Outline

- SINTEF
- Performance in Discrete Optimization
- Hardware developments, and prospects
- Accelerators and heterogeneous computing
- «Camel Spider (Solifugae)» - a GPU based VRP solver
- Extension to truly heterogeneous computing
- Conclusions

Stream processing accelerators

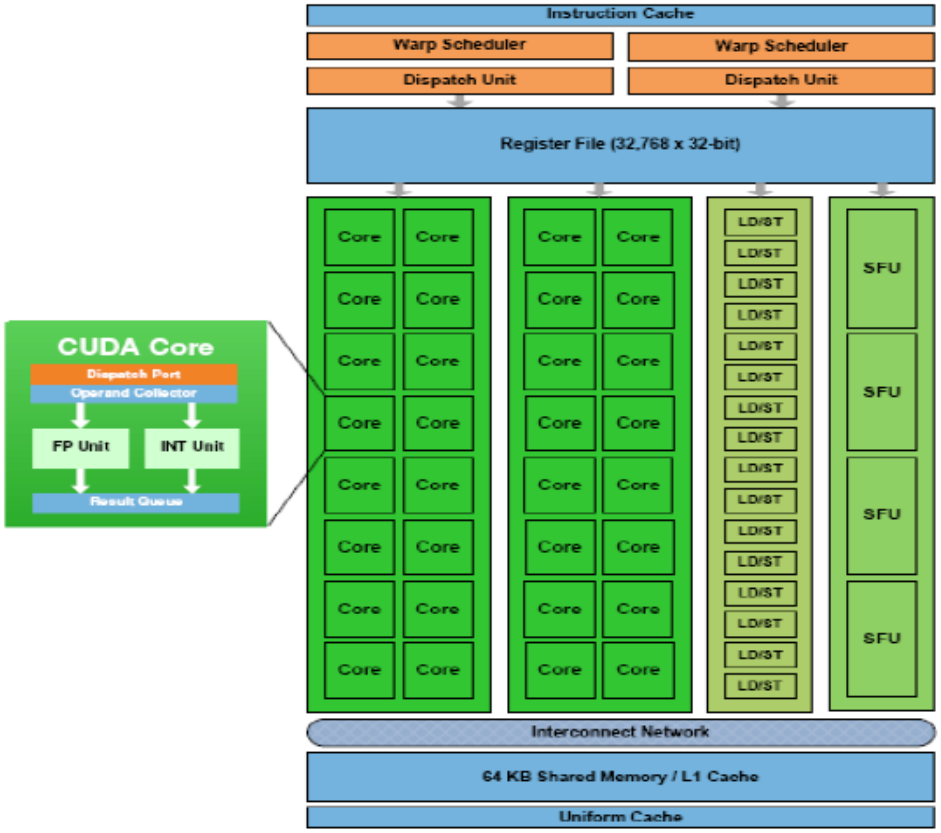
- The graphics card was the origin
- Developent driven by gaming industry
- Computing power increases rapidly
- Programmability improves rapidly
- Libraries, debugging and profiling tools
- Single Program Multiple Data
- Massively parallel, thousands of threads
- You need to
  - understand the architecture
  - worry about code diversion
  - worry about memory latency

The GPU – NVIDIA Fermi Architecture



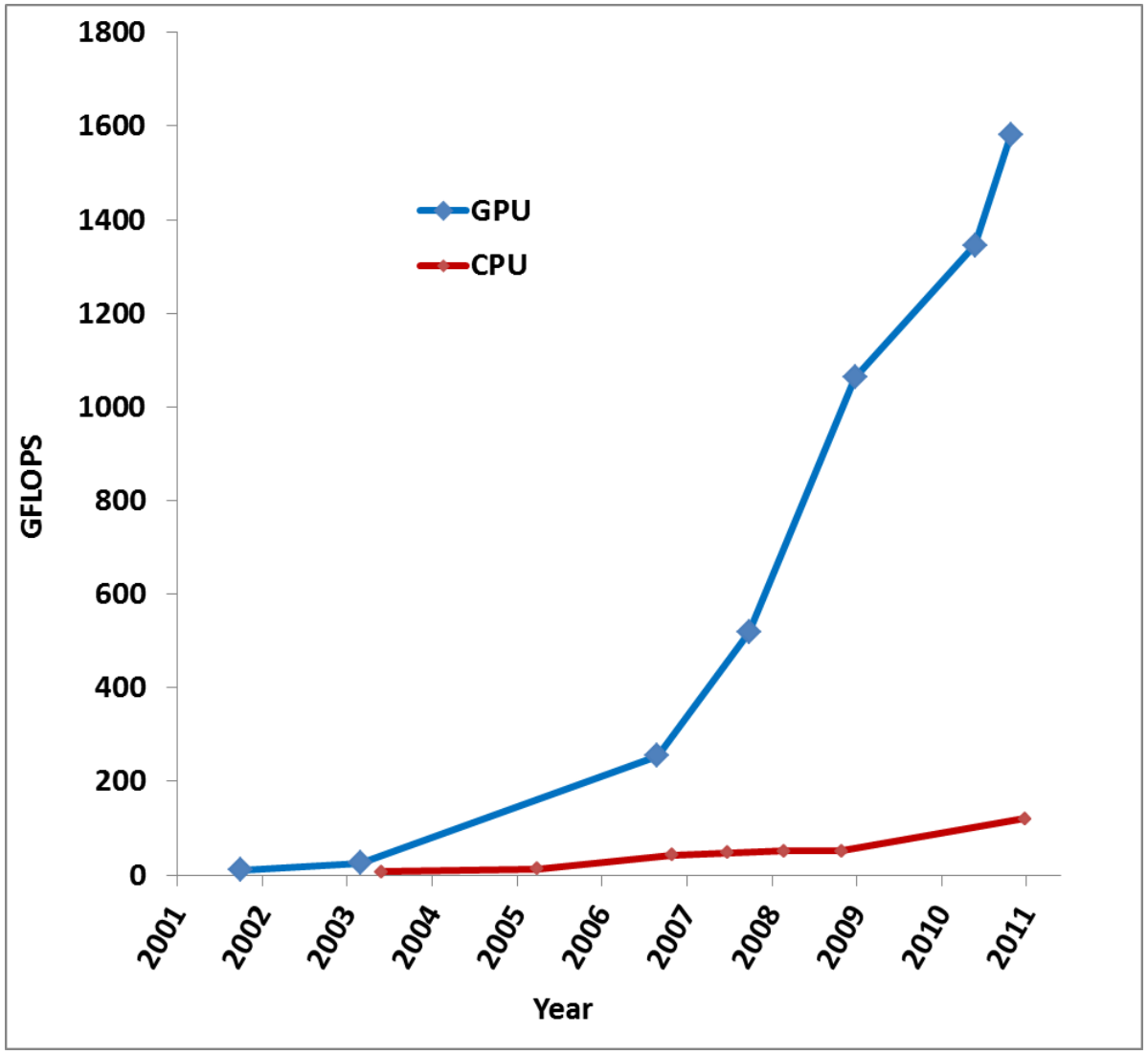
16 streaming multiprocessors are positioned around a common L2 cache

The GPU – NVIDIA Fermi Architecture



Each of the16 Streaming Multiprocessors (SMs) has 32 cores, 512 cores in total. Each core run the same program («kernel»), with individual data and individual code flow (SPMD). Divergence means serialization. Need more threads than cores to hide latency, typically >512 threads for each SM, say 10.000. One may run multiple kernels concurrently.

GPU vs CPU performance



Programming GPUs

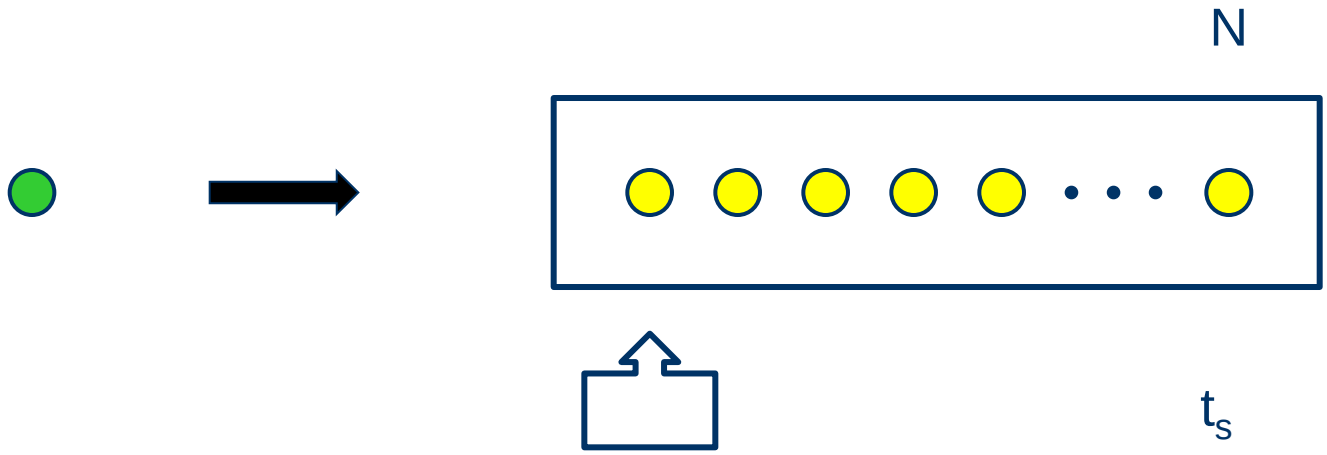
- CUDA
  - C++-like language
  - proprietary (NVIDIA)
  - libraries
  - development tools
- OpenMP
  - API for multi-platform shared memory multiprocessing
  - C, C++, Fortran
  - Open standard, Khronos group

Exploiting the GPU

- Games
- Matrix and vector operations
- Scientific simulation and visualization http://www.youtube.com/babrodtk
- http://www.nvidia.co.uk/object/cuda_apps_flash_new_uk.html#

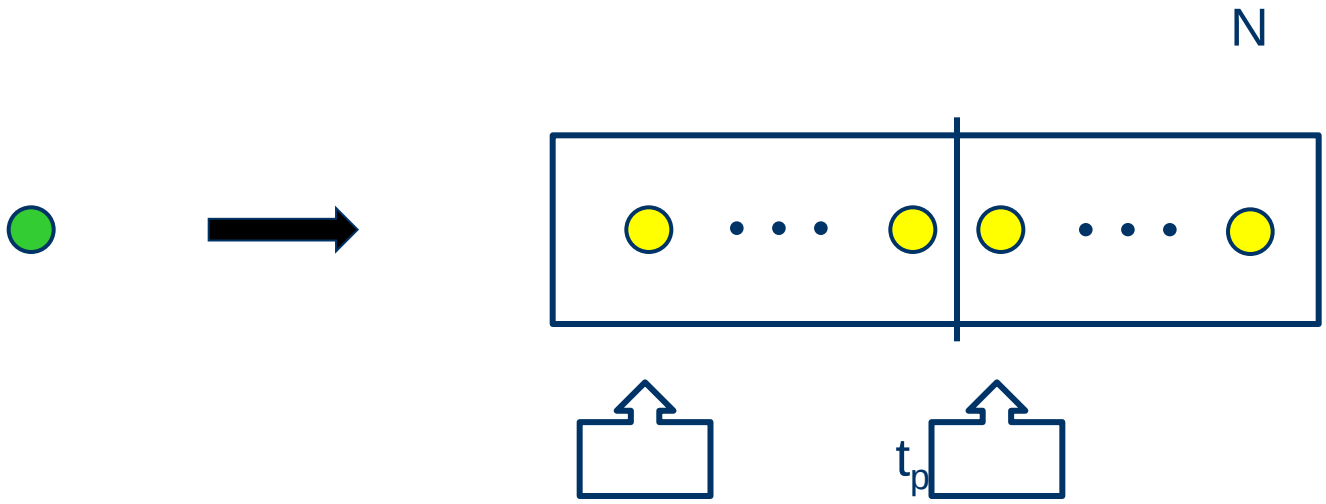
- Local search
- Genetic algorithms

Local Search
- Sequential evaluation of neighborhood



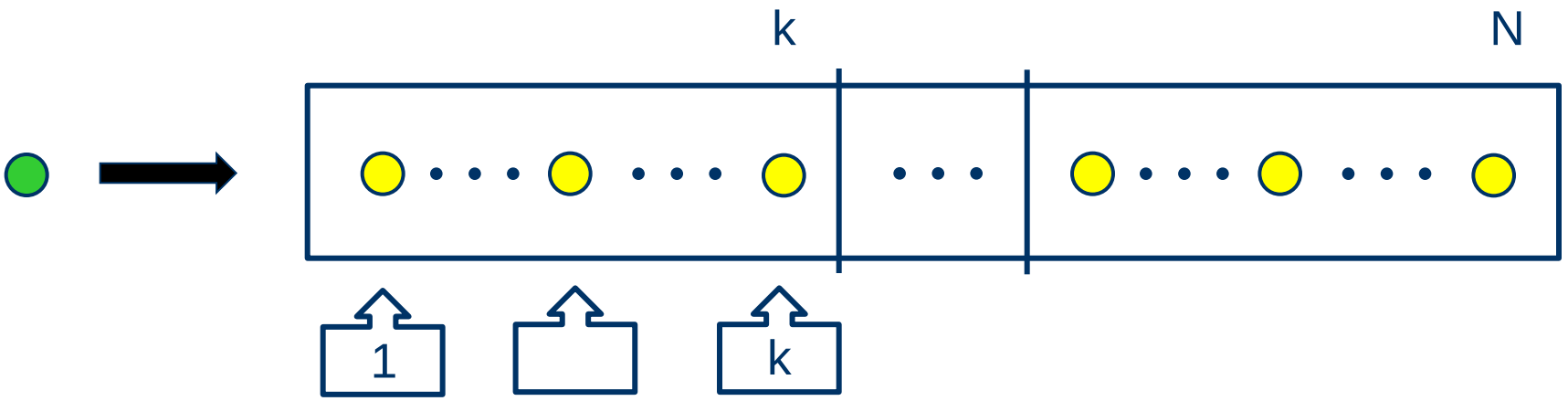
Time for one iteration: $t_s N$

Local Search
– Task parallel evaluation of neighborhood 2 cores



Time for one iteration: $t_pN/2$

Local Search
– Data parallel evaluation of neighborhood



simultaneous kernels: k

Time per evaluation: t_g

Time for one iteration: $t_g N/k$

Heterogeneous computing

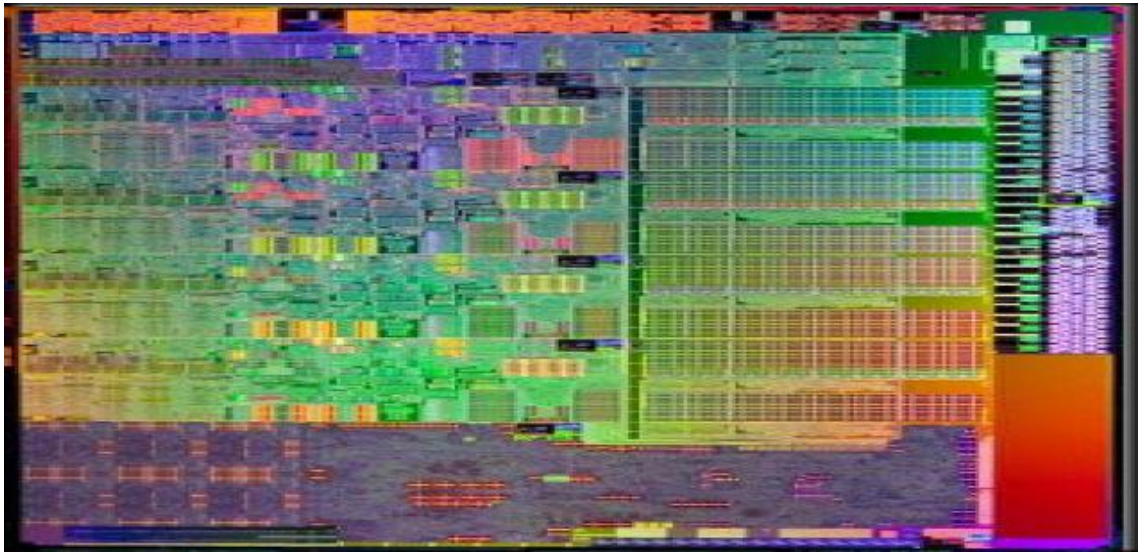
- **Heterogeneous computing systems**:
  electronic systems that use a variety of different types of computational units.
- Current and future PCs are parallel and heterogeneous

"GPUs have evolved to the point where many real-world applications are easily implemented on them and run significantly faster than on multi-core systems. Future computing architectures will be hybrid systems with parallel-core GPUs working in tandem with multi-core CPUs."

Prof. Jack Dongarra
Director of the Innovative Computing Laboratory
The University of Tennessee

Supercomputer on a chip
Single die heterogeneous processors

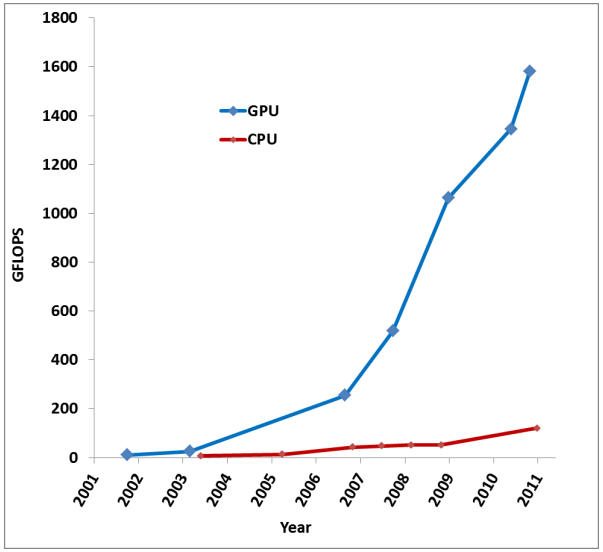
- AMD Fusion
- Intel Sandy Bridge



Why bother?

- Exploit present hardware
- Profit from the future increase of processor power

- Robustness
- Larger-size, more integrated problems
- Real-time applications
- Stochastic models
- Multi-criteria problems
- New ideas in optimization
- Automated parallelization?
- Tool vendors?



Literature

- Van Luong T.V., Melab N., Talbi E.-G.: **Neighborhood Structures for GPU-based Local Search Algorithms.** Parallel Processing Letters, Vol. 20, No. 4, pp. 307-324, December 2010
- Hasle G., Kloster O., Riise A., Schulz C., Smedsrud M.: **Using Heterogeneous Computing for Solving Vehicle Routing Problems**. Extended abstracts TRISTAN VII, Tromsø, Norway, June 20-25, 2010.
- Schulz C., Hasle G., Kloster O., Riise A., Smedsrud M.: **Parallel local search for the CVRP on the GPU.** META'10, Djerba, Tunisia, October 28 2010
- Special session «**Metaheuristics on graphics hardware**» at META'2010 http://www2.lifl.fr/META10/pmwiki.php?n=Main.InfoMGH
- Call for papers JPDC Special Issue: Metaheuristics on GPU

Outline

- SINTEF
- Performance in Discrete Optimization
- Hardware developments, and prospects
- Accelerators and heterogeneous computing
- «Camel Spider (Solifugae)» - a GPU based VRP solver
- Extension to truly heterogeneous computing
- Conclusions

Activities at SINTEF

- PDA-based simulation, geometry, visualization
- Collab project 2009-2012
  - INRIA, University of Antwerp, CIRRELT, Olli Bräysy, SINTEF
- Task parallelization of the industrial VRP Solver «Spider»
- Experimental VRP solver: «Camel Spider»
- Project workshops
- META'2010 special session
- JPDC special issue

Camel Spider

- Experimental VRP solver, heterogeneous computing
- Giant Tour Representation
- Resource Extension Functions (REFs)
- Segment hierarchy for constant time evaluation
- Irnich S.: “A Unified Modeling and Solution Framework for Vehicle Routing and Local Search-based Metaheuristics” INFORMS JOURNAL ON COMPUTING, Vol. 20, No. 2, Spring 2008, pp. 270-287

- How efficient can we make local search using the GPU?

Camel Spider

- Capacitated (Distance constrained) VRP

- CPU: Send problem data to GPU
- GPU tasks
  - Create neighborhood
  - Each iteration
    - create segment hierarchy
    - evaluate capacity constraint, # tours, length constraint, length objective
    - select best move (and send information to CPU)

Camel Spider – C/DVRP Experiments

- Local Search with 2-opt, 3-opt on Giant Tour
- NVIDIA GTX480 (Fermi architecture)
- Benchmarks 30-1200 customers
- Li, Golden, Wasil 1200.vrp*
  - 1.2 billion 3-opt moves generated and evaluated in 14 s (12 ns per move).
  - GPU occupancy >60%
- Speedup factor vs. serial CPU up to almost 1000

- Average solution quality poor ...

*Li, F., B.L. Golden and E.A. Wasil (2005). Very large-scale vehicle routing: New test problems, algorithms and results. Computers & Operations Research 32, 1165–1179.

Camel Spider – Lessons GPU for LS

- GPU has great potential for speedup of Local Search
- Fine tuning may contribute with factor 5-10
- Relatively simple, large neighborhoods should be targeted
- Neighborhood must be large enough for evaluation to dominate
- 2-opt only for large instances

Camel Spider – 2-opt, 575 nodes

Camel Spider – 2-opt, 2400 nodes

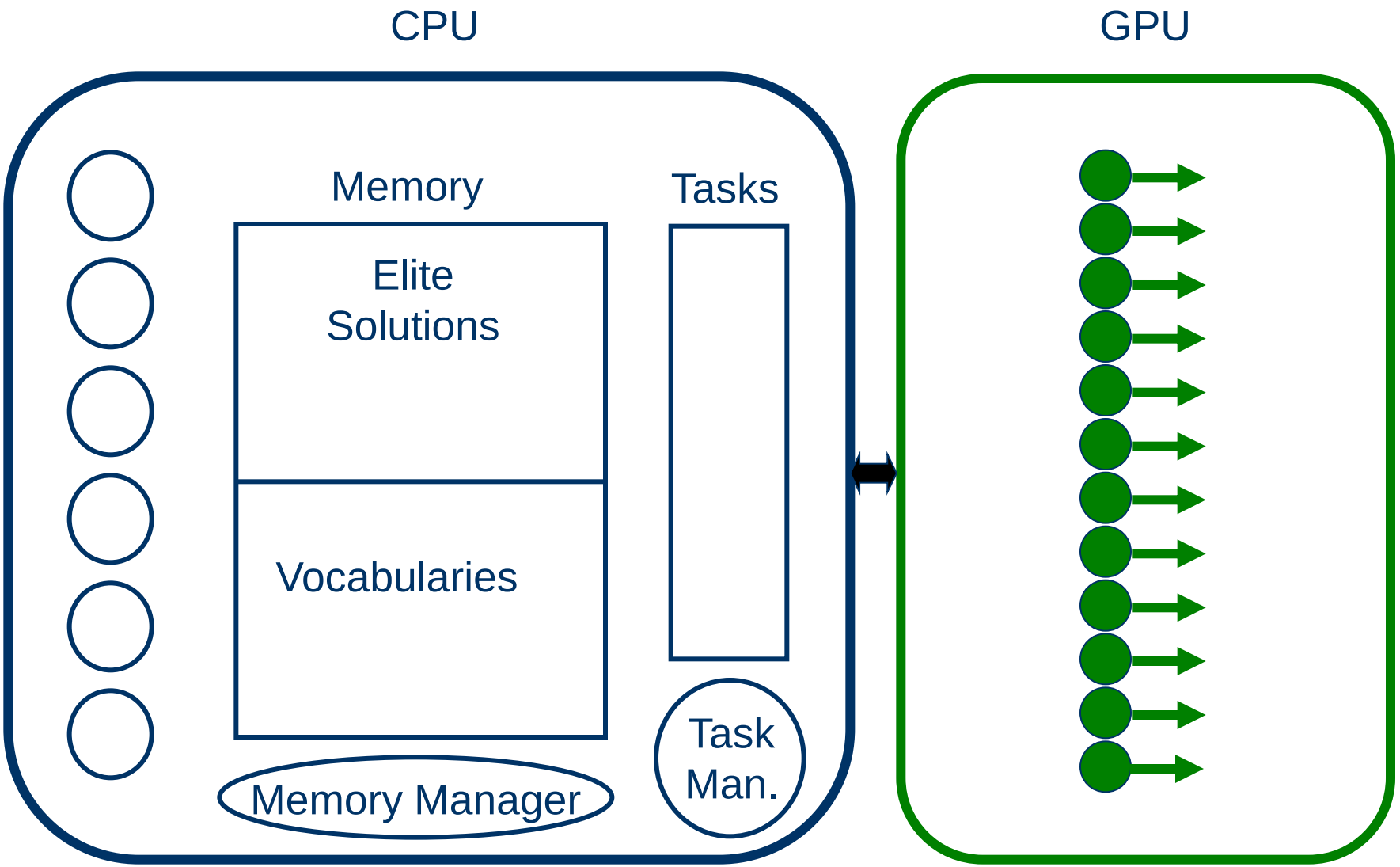
Outline

- SINTEF
- Performance in Discrete Optimization
- Hardware developments, and prospects
- Accelerators and heterogeneous computing
- «Camel Spider (Solifugae)» - a GPU based VRP solver
- Extension to truly heterogeneous computing
- Conclusions

Ideas – Heterogeneous DOP Computing

- Goal: Balanced use of available computing devices
- Self-adaptation to available hardware
- The GPU is a good intensification machine
  - Local Search
  - Large Neighborhood Search
  - Variable Neighborhood Search
  - ...
- ... but one needs to worry about code diversion and memory management
- CPU used for more «sophisticated» tasks

Sketch of labor division – VRP Solver



Questions

- Exact methods?

- How to assess the performance of an optimization algorithm?
- What is «within reasonable time»?

Conclusions

- «The Beach law» does not hold any more ...
- Fundamental change in the general increase of computing power
- «Moore's law» still at work, until 2030?
- The GPU has become a generally programmable, very powerful device
- Local search up to 1000 times faster on the GPU than on one CPU core
- Every PC will soon have a heterogeneous supercomputer inside
  - multiple cores
  - stream processing accelerator
- Your sequential program can only exploit a small fraction of the power
- Little hope of efficient tools for automatic parallelization
- Providers of basic optimization technology cannot ignore the potential
- Bottlenecks in industry and research
- Opportunities for new ideas in optimization

“The Beach Law” does not hold any more

Discrete optimization needs heterogeneous computing

Christian Schulz, Trond Hagen, Geir Hasle
Department of Applied Mathematics, SINTEF ICT, Oslo, Norway

Seminar

CORAL, Aarhus School of Business and Social Sciences
Aarhus, Denmark, March 16, 2011