ENERGIE
GESUNDHEIT
INFORMATION
UMWELT

Aspects of computer architectures

Lattice Practices '06

27-28 November 2006

Norbert Eicker
John von Neumann Institute for Computing

NIC

Forschungszentrum Jülich
in der Helmholtz-Gemeinschaft

50 Jahre Zukunft

Outline

ENERGIE GESUNDHEIT INFORMATION UMWELT

- Motivation
- Lattice QCD Building blocks
- General Architecture
- Processor Aspects
- Memory Bandwidth Limitations
- Network Constraints
- Summary

Motivation

ENERGIE GESUNDHEIT INFORMATION UMWELT

- Lattice QCD is a powerful tool deviate predictions from first principles of QCD
- Almost all results depend on statistical methods
- Big and complex numerical problem
  - Runtime measured in years not unusual
- Need of big strong computers
- Optimization inevitable
- Optimizing beyond certain level still needs deeper understanding of computer-architecture

LQCD Building blocks

- Computer hardware only supports floating-point operations (kind of real numbers)
- LQCD depends on complex operations:
  - Complex MUL: 4 MUL + 2 ADD
  - Complex ADD: 2 ADD

$$\vec{y} = \alpha \vec{x} + \vec{y}$$

- Complex triad:
  - Operations: (4 MUL + 2 ADD + 2 ADD) x N
  - Load/Store: (2 x 2 LOAD + 2 STORE) x prec x N

LQCD Building blocks

ENERGIE GESUNDHEIT INFORMATION UMWELT

- SU(3) x color:

$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \begin{pmatrix} A \\ B \\ C \end{pmatrix} = \begin{pmatrix} aA+bB+cC \\ dA+eB+fC \\ gA+hB+iC \end{pmatrix}$$

  - Operations: 36 MUL + 18 ADD + 12 ADD
  - Load/Store: 2x(9+3) x prec LOAD
    2x3 x prec STORE

LQCD Building blocks

ENERGIE GESUNDHEIT INFORMATION UMWELT

- Wilson kernel: $(1-\gamma_\mu)U_\mu(x)\phi(x+\hat{\mu})$
  - Operations: 2 x [36 MUL + (30 + 6) ADD]
  - Load/Store: 2 x (9+12) x prec LOAD
    2x12 x prec STORE
- Wilson matrix:

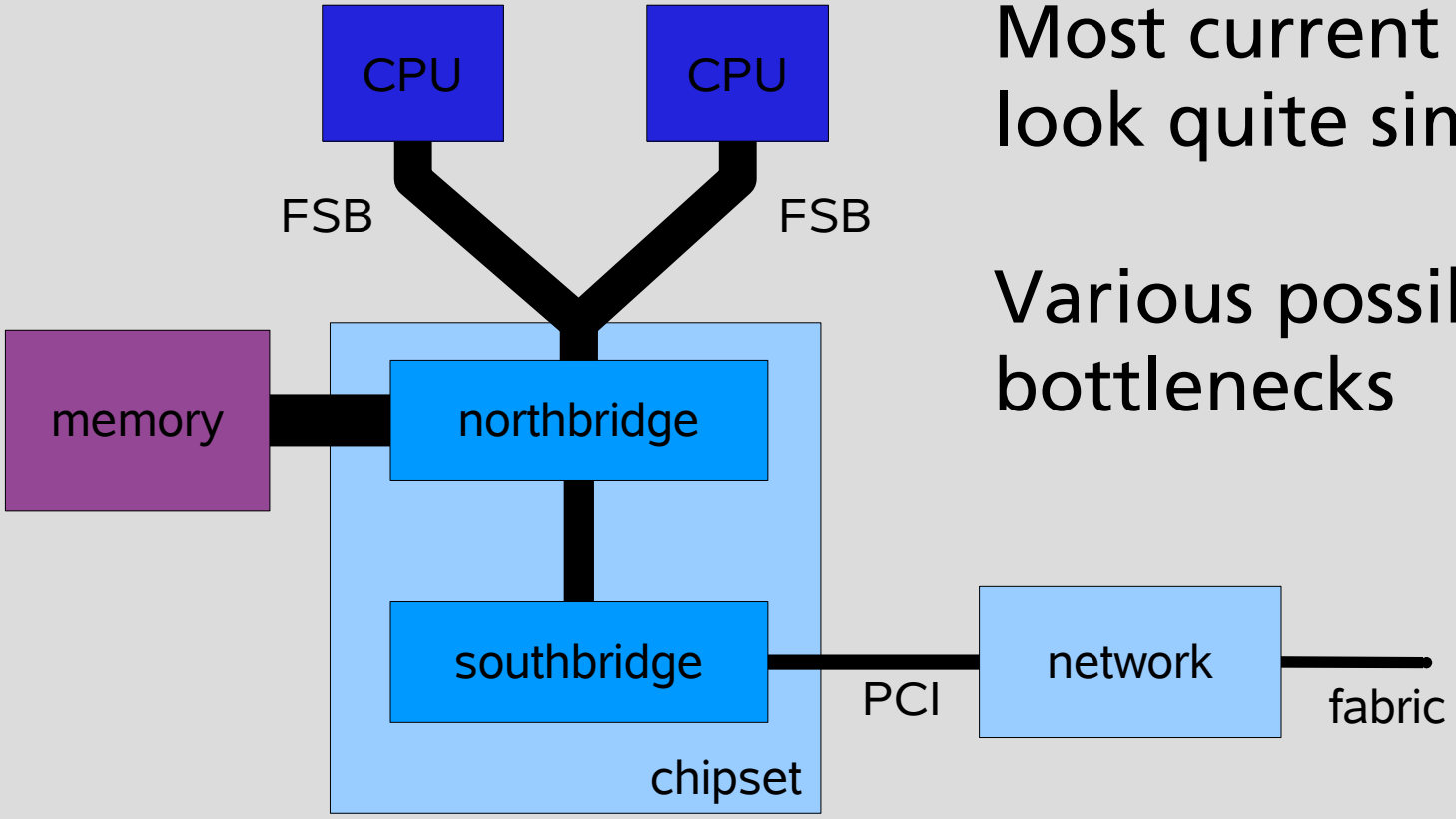
$$\phi(x)-\kappa\sum\left[(1-\gamma_\mu)U_\mu(x)\phi(x+\hat{\mu})+(1+\gamma_\mu)U_\mu^+(x-\hat{\mu})\phi(x-\hat{\mu})\right]$$

  - Operations: 8 x 72 + 2 x 12 = 600 MUL
    8 x 72 + 7 x 2 x 12 + 2 x 12 = 768 ADD
  - Load/Store: (4 x 2 x 9 + 2 x 12) = 96 x prec LOAD
    2 x 12 = 24 x prec STORE

Definitions

- Bandwidth / Throughput:
  - Number of operations / bytes per unit of time
    - Memory: ~ 10 Gbyte/sec (or more from cache)
    - Network: ~ 1 GByte/sec
    - Processor: ~ 10 GFlops
- Latency:
  - Time needed for startup of transmission/operation
    - Memory: ~ 1 µsec
    - Network: 2-20 µsec
    - Processor: ~100 cycles (or less from cache)

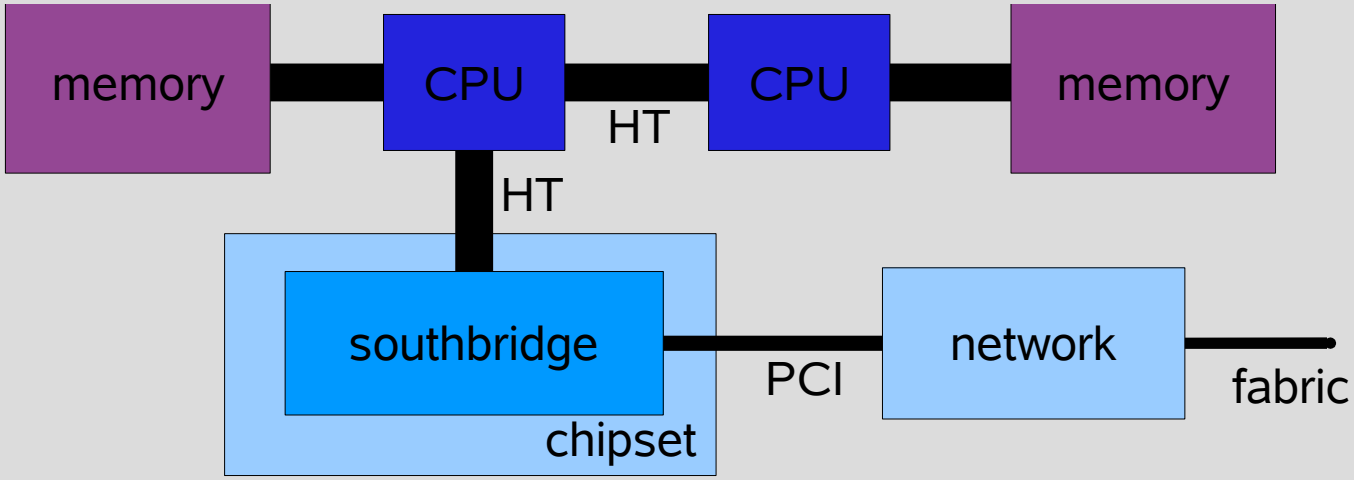
General Architecture



Most current systems look quite similar

Various possible bottlenecks

General Architecture



- AMD current advantage
- Allows to build scalable SMPs
- For Intel's architecture early 2008:
  - Replace HT with CSI

Processor Aspects

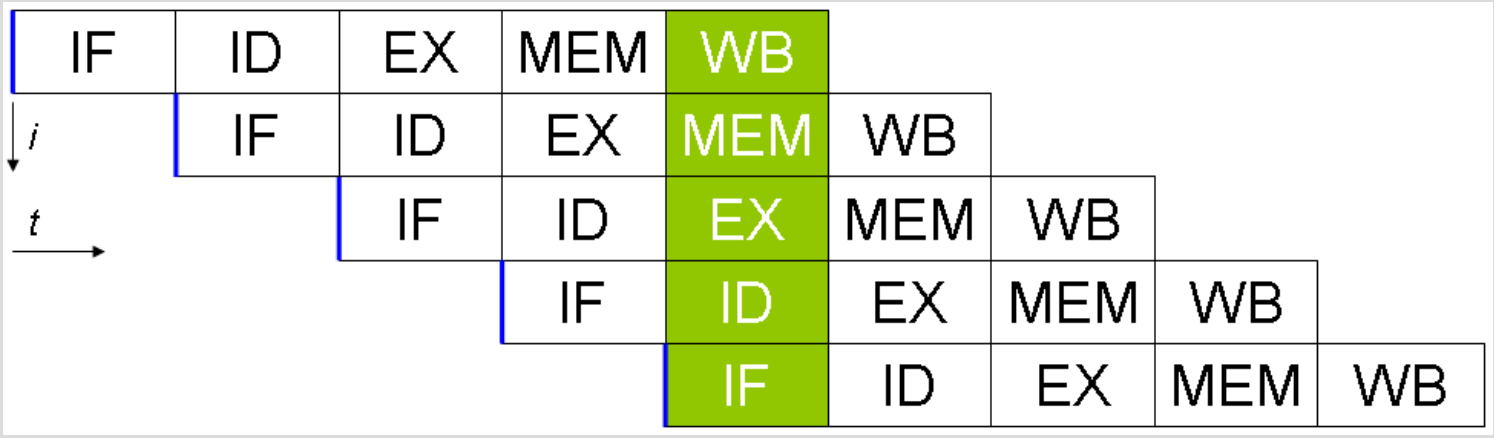
ENERGIE GESUNDHEIT INFORMATION UMWELT

- Almost all current CPU architectures are

  Superscalar pipelined RISC

- Superscalar: More than one operation per cycle
  - 1 or 2 FPU for MULT and ADD operations
  - 4+ Integer Units (also handling LOAD and STORE)
- Pipelined
  - Issue one operation per cycle
  - Actual operation takes some cycles (4-8)
- RISC: Reduced Instruction Set → higher clock

Pipelining



- Latency (few cycles) to fill pipeline
- Usually DIV, SQRT break pipeline (avoid!)
- Length of pipeline determines availability of result
- Register might still be blocked (register files)

Operation Scheduling

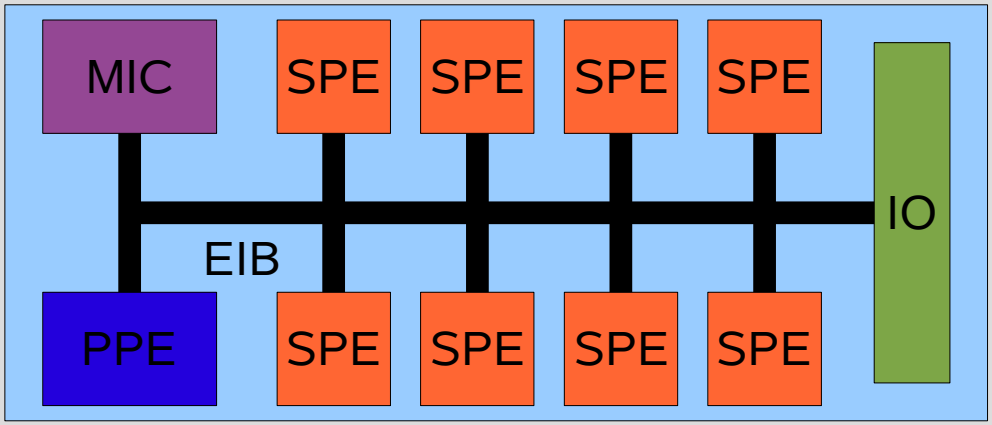
- Operation's result is delayed due to pipeline
- Next one might depend on yet unavailable results
- Out of order execution
  - Delay next operation until all operands available
    - Pro: Easy to write code & compilers
    - Con: Complex processor-design
- VLIW – Very Long Instruction Word
  - Scheduling hard code in Instruction words
    - Pro: Leaner processor design → Higher clocks
    - Con: More complex compiler, cache problems

Multimedia Extensions

ENERGIE GESUNDHEIT INFORMATION UMWELT

- Modern CPUs support Vector Operations
  - Similar to former vector architecture
  - Unfortunately without memory bandwidth
  - Multiple concurrent operations in wide registers
- MMX, SSE, SSE2, SSE3, ...
- Usually only single precision
- Capability to handle complex arithmetic varies
  - Typically no explicit complex operations
  - Intra-Register swap for CMULT needed

Special Architecture - Cell



- PPE: Dual-Core PowerPC w/ 512 kB cache
- SPE: FPU + 256 kB memory (> 200 GFlop SP)
- EIB: up to 96 Bytes per cycle
- SDK and simulator available from IBM

Special Architecture

ENERGIE GESUNDHEIT INFORMATION UMWELT

- GPU
  - Todays Graphics Hardware provides significant compute-power & memory bandwidth
    - Usually only single precision (IEEE conform?)
    - SDK available
- FPGA
  - Adapt the compute hardware to your problems
    - Hard to program (VHDL)
    - Memory bottlenecks
    - How to attach the network

Performance - CPU

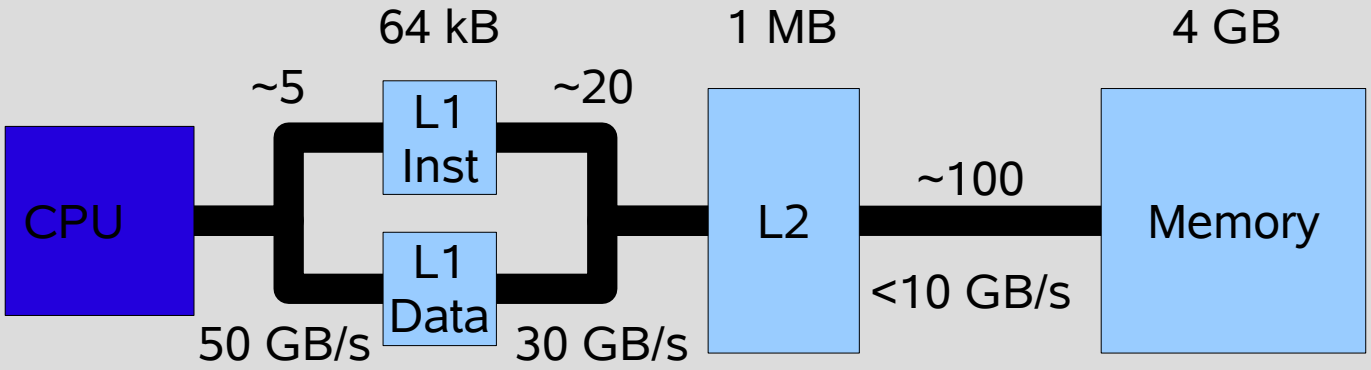
ENERGIE GESUNDHEIT INFORMATION UMWELT

- Complex Triad:
  - ADD and MULT operations are balanced (4 + 4)
  - Latency negligible for long vectors
  - Possible to get almost peak performance
- Wilson Matrix
  - More ADD (768) than MULT (600) operations
  - There will be empty MULT slots
  - Not more than ~90% possible
  - Number of register vs. pipeline stages

Memory

- Main memory is slow (compared to CPU)
- Will become worse in future
  - Different growth-rates for mem & CPU clock
  - Dual-Core CPU do not double bandwidth
  - Multi-Core CPU
- Problems
  - Bandwidth cannot sustain CPU for triad
  - Latency of main memory ~100 cycles

Caches



- Hide slow memory behind caches
  - More layers available
  - Some layers might be shared between CPUs
  - Different levels of associativity
  - Network might be attached to Cache!

Caches

- Caches typically hard to control
  - No control what stays in cache
  - Smart heuristic might fail for given problem
- Pre-fetching
  - Get data into distinct level of cache before use
  - Try to hide memory latency
- Communication might spoil caches!

Performance - Memory

ENERGIE GESUNDHEIT INFORMATION UMWELT

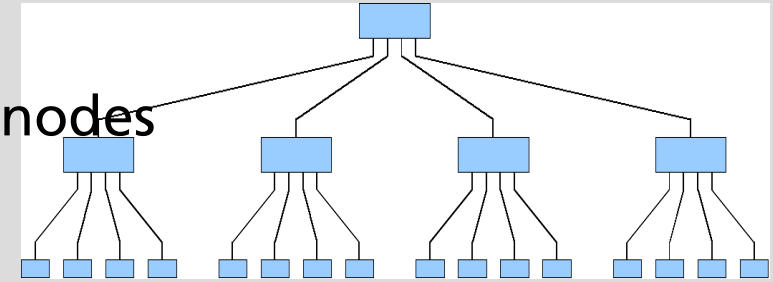
- Triad (example 3 GHz, 1 MULT/ADD FPU, 6 GB/s):
  - In L1 cache: peak (i.e. 4 GFlops)
  - Bandwidth needed:
    - 4 MUL + 4 ADD → 1 vector entry / 4 cycles
    - 4 LOAD + 2 STORE / 4 cycles
      - Single precision: 4 Bytes x 6 / 4 cycles = 18 GB/s
      - Double precision: 8 Bytes x 6 / 4 cycles = 36 GB/s
  - Effective performance:
    - Single precision: 2 GFlops (i.e. 33% peak)
    - Double precision: 1 GFlops (i.e. 17% peak)

Performance - Memory

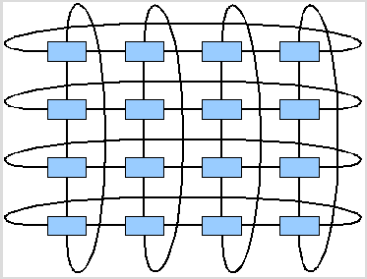
- Wilson kernel (3 GHz, 1 MULT/ADD FPU, 6 GB/s)
  - Operations: 600 MUL + 768 ADD / 768 cycles
  - Load/Store: 96 x prec LD + 24 x prec ST / 768 cycles
    - Single precision: 480 Byte / 768 cycles = 1.875 GB/s
    - Double precision: 960 Byte / 768 cycles = 3.75 GB/s
- Effective performance not spoiled by Bandwidth!
- But:
    - Cache misses might lead to further LDs
    - Effective Bandwidth typically smaller
    - Solver also contains triad parts

Network Topologies

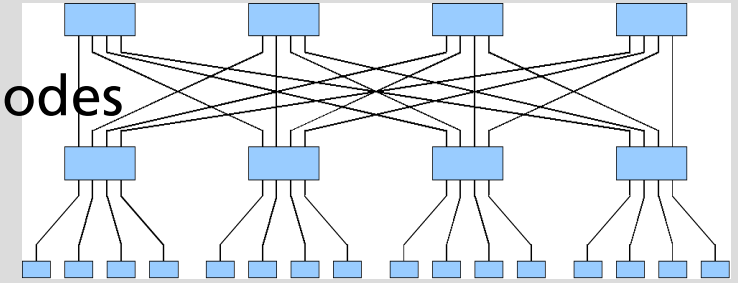
- Fat Tree
  - Communication between arbitrary nodes
  - Varying distances
  - Bottlenecks



- Torus / Mesh
  - Only nearest neighbor communication
  - Cut through routing possible
  - Scalable for adequate applications



- Clos / ω-Network
  - Communication between arbitrary nodes
  - Varying distances
  - No contention – at least in principle



Network

- LQCD only needs nearest neighbor communication
  - But we have to deal with global sums
- Messages in LQCD are relatively small
  - Assume local lattice with 4^4 sites:
    - For SU(3): 4^3 x 4 x 2 x 9 x prec = 4608 x prec
    - For color-spinor: 4^3 x 3 x 4 x 2 x prec = 1536 x prec
    - gamma-projection gives another factor of 2
- Quite sensitive to network latencies
  - 2 µsec means more than 4000 cycles!

Network

ENERGIE GESUNDHEIT INFORMATION UMWELT

- Crucial parameters:
  - Bandwidth & Latency (as always)
  - Packet rate: How many (small) messages can be sent
  - CPU usage: Is the main CPU free for computation
  - Cache Coherence: Does communication spoil the caches
- Try to hide communication as much as possible behind computation (asynchronous)
- Latency is most crucial

Performance - Network

- Triad:
  - Quite easy: No communication, no effect :-)
- Wilson matrix (local lattice with 4^4 sites):
  - 3 GHz, 1 MULT/ADD FPU, 6 GB/s, 2 µsec, 1 GB/s
  - 2 µsec = 6000 cycles = 8 sites
    - Single precision: 1536 x 4 B / 1 GB/s = 6.1 µsec
    - Double precision: 1536 x 8 B / 1 GB/s = 12.3 µsec
  - 8 directions x 12.3 µsec = 384 sites (> 4^4 = 256)
  - Even worse due to latency, caches, SMP, etc.

Summary

ENERGIE GESUNDHEIT INFORMATION UMWELT

- LQCD needs maximum performance
- Performance optimization needs insight into computer architecture
- Optimal performance has to consider every layer in system architecture
- Pipelined Processors need fine-tuned code
- Many layers of caches enforce pre-fetching and cache-reuse
- Limited scalability due to sensitivity on latency