Node-Level Performance Engineering

For final slides and example code see:
https://tiny.cc/NLPE-SC20

Georg Hager, Jan Eitzinger, Gerhard Wellein
Erlangen Regional Computing Center (RRZE)
and Department of Computer Science
University of Erlangen-Nuremberg

SC20 full-day tutorial
November 9-10, 2020
Atlanta, GA

Agenda

- Part I
  - Introduction to compute node architecture
  - Performance tools 1: topology and affinity
  - Microbenchmarking as a tool
  - Demo
  - Introduction to the Roofline model
  - Performance tools 2: hardware performance counters
  - Demo

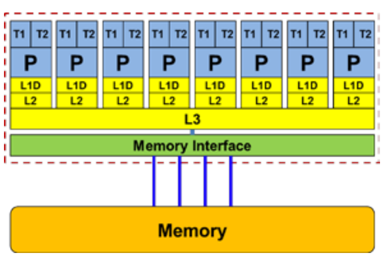
- Part II
  - Case study: tall & skinny matrix-matrix multiplication
  - Case study: Stencil codes
  - Demo
  - Case study: sparse matrix-vector multiplication
  - Programming for Single Instruction Multiple Data (SIMD) parallelism
  - Programming for ccNUMA

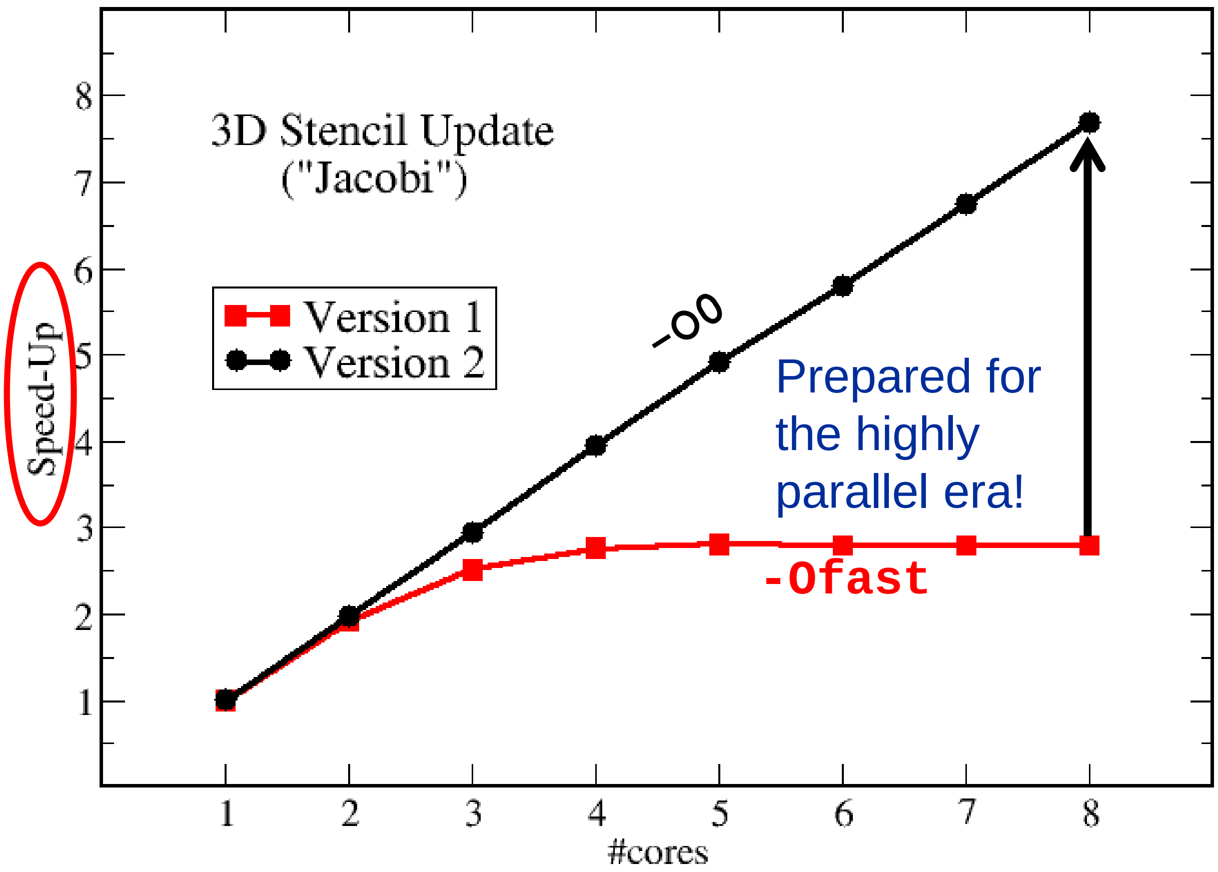
Prelude:
Scalability 4 the win!

Scalability Myth: Code scalability is the key issue

```
!$OMP PARALLEL DO
do k = 1 , Nk
  do j = 1 , Nj; do i = 1 , Ni
     y(i,j,k)= b*(  x(i-1,j,k)+ x(i+1,j,k)+ x(i,j-1,k)+
                    x(i,j+1,k)+ x(i,j,k-1)+ x(i,j,k+1))
    enddo; enddo
enddo
!$OMP END PARALLEL DO
```

Changing only the compile options makes this code scalable on an 8-core chip

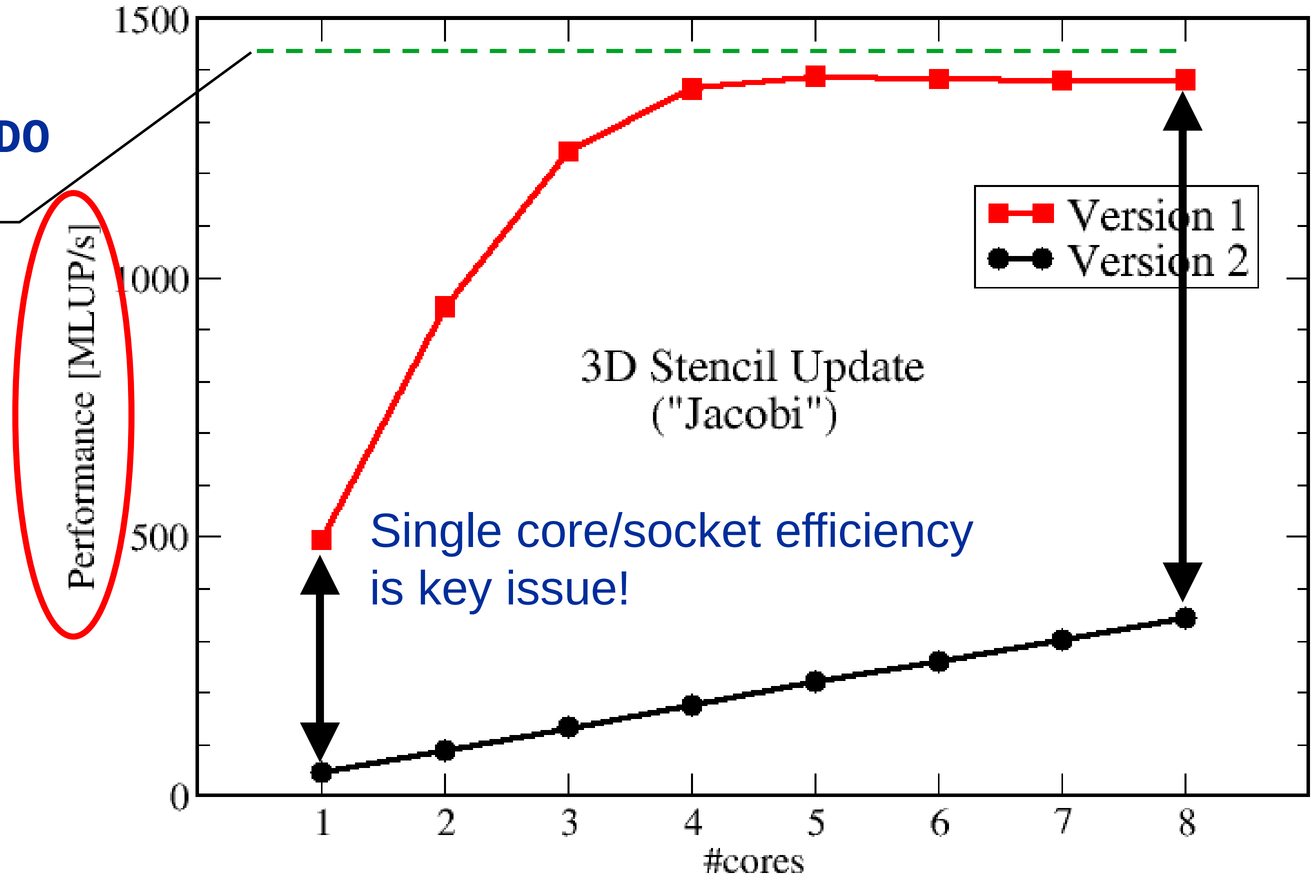


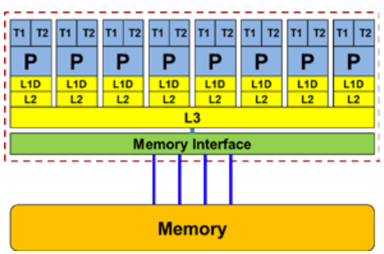


Scalability Myth: Code scalability is the key issue

```
!$OMP PARALLEL DO
do k = 1 , Nk
  do j = 1 , Nj; do i = 1 , Ni
     y(i,j,k)= b*(  x(i-1,j,k)+ x(i+1,j,k)+ x(i,j-1,k)+
                    x(i,j+1,k)+ x(i,j,k-1)+ x(i,j,k+1))
    enddo; enddo
enddo
!$OMP END PARALLEL DO
```

Upper limit from simple performance model:
35 GB/s & 24 Byte/update





Questions to ask in high performance computing

- Do I understand the performance behavior of my code?
  - Does the performance match a model I have made?

- What is the optimal performance for my code on a given machine?
  - High Performance Computing == Computing at the bottleneck

- Can I change my code so that the “optimal performance” gets higher?
  - Circumventing/ameliorating the impact of the bottleneck

- My model does not work – what’s wrong?
  - This is the good case, because you learn something
  - Performance monitoring / microbenchmarking may help clear up the situation