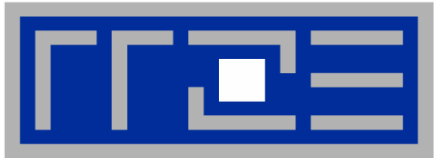
Effiziente Nutzung von Hochleistungsrechnern in der numerischen Strömungsmechanik

Dr. Georg Hager
georg.hager@rrze.uni-erlangen.de

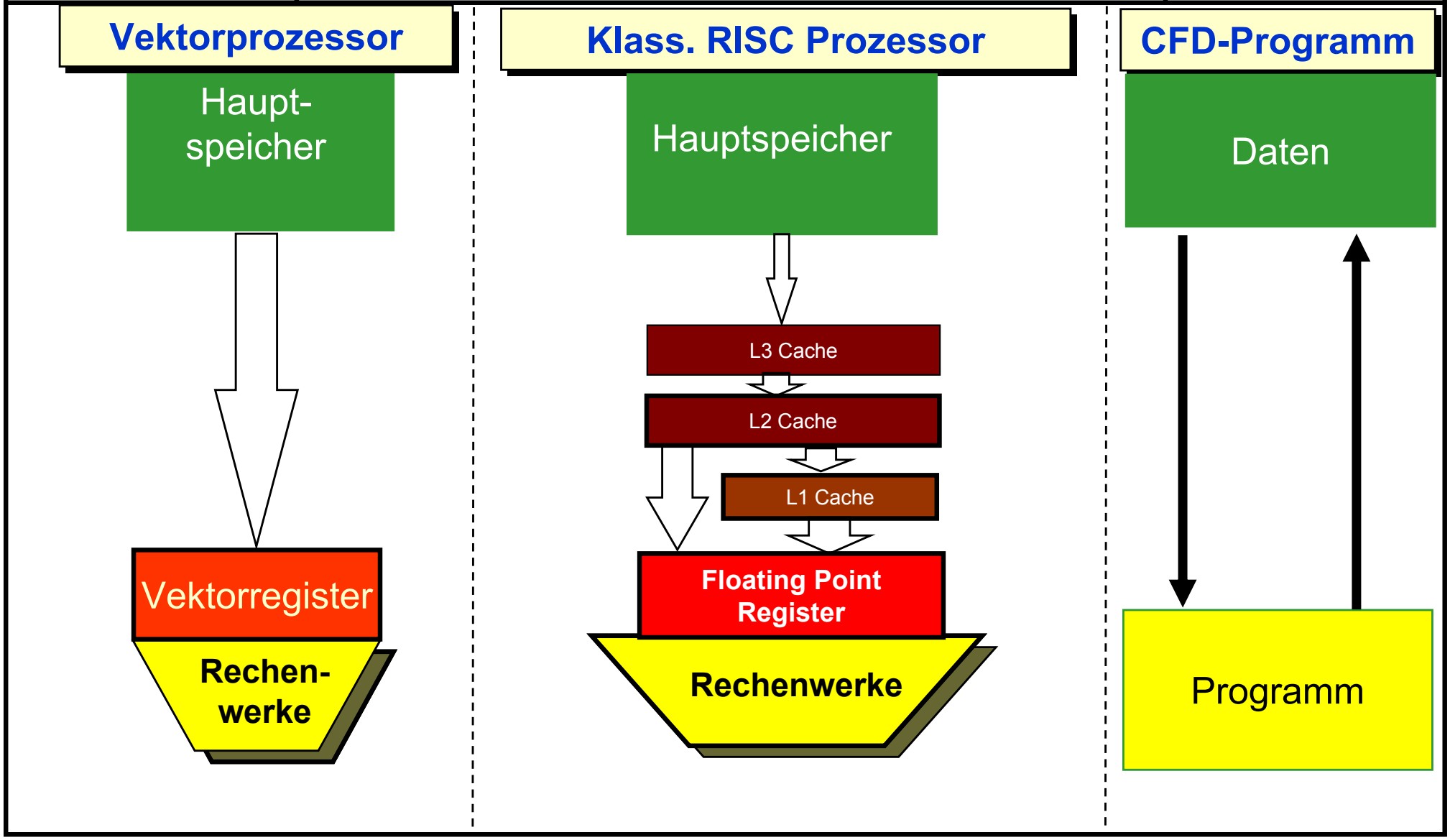
HPC Services
Regionales RechenZentrum Erlangen



- Moderne Prozessoren für numerische Anwendungen
  - Allgemeine Optimierungsrichtlinien für CFD
- Parallelrechner
  - Möglichkeiten und Grenzen der Parallelität
  - Designprinzipien
  - effiziente Programmierung für CFD-Probleme
- Beispiel: Performance und Parallelisierung des SIPSolvers nach Stone (Finite-Volumen)
- Zusammenfassung

Moderne Prozessoren für numerische Anwendungen

Speicherhierarchien moderner Prozessoren
Schematischer Aufbau



- **Rechenleistung** in Fließkommaoperationen pro Sekunde: **Flop/s**
  - i.A. Multiplikationen und Additionen mit doppelter Genauigkeit
  - Divisionen, SQRT etc. sind oft **sehr langsam**
  - max. FLOP/s-Zahl ist üblicherweise das Doppelte oder Vierfache der Prozessor-Taktfrequenz
- **Bandbreite** (BW), d.h. Geschwindigkeit des Datentransfers: **GB/s**
  - in CFD-Anwendungen meist der limitierende Faktor: Bandbreite zwischen CPU und Hauptspeicher
- **Latenz**: Zeitdauer vom Anstoßen eines Datentransfers bis zum Eintreffen des ersten Bytes
  - CPU↔CPU, CPU↔Speicher, Rechenwerke↔Cache
  - dominant, wenn pro Transfer nur wenig Daten übertragen werden

Kurzlehrgang NUMET	Typische Leistungszahlen moderner Prozessoren	6 / 11

		Intel Xeon	AMD Opteron	Intel Itanium2	NEC SX8
Spitzenleistung Taktfreq.		12 GFlop/s 3.0 GHz	8 GFlop/s 2.0 GHz	6.4 GFlop/s 1.6 GHz	16 GFlop/s 2.0 GHz
# FP Registers		16/32	16/32	128	8 x 256 (Vector)
L1	Size	32 kB	64 KB	16 KB (kein FP)	---
	BW	96 GB/s	64 GB/s	32 GB/s	---
	Latenz	3 Takte	3 Takte	1 Takt	---
L2	Size	4 MB (shared)	0.5 MB	256 KB	---
	BW	48 GB/s (sh.)	51.2 GB/s	51.2 GB/s	---
	Latenz	14 Takte	12 Takte	5-6 Takte	---
L3	Size	---	2 MB (shared)	9 MB	---
	BW	---	???	51.2 GB/s	---
	Latenz	---	???	14 Takte	---
Mem.	Socket BW	10.6 GB/s r/w	10.6 GB/s r/w	8.5 GB/s r/w	64 GB/s r/w
	Latenz	~150 ns	~80 ns	~200 ns	Vektorisierung

- Caches bestehen aus **Cachelines**, die nur als Ganzes gelesen/geschrieben werden
  - typische Längen: 8/16/32 Worte (je 8 Byte)
- **Cache Miss**: Verwenden eines Datums, das noch nicht im Cache liegt
  - Folge: Cacheline wird aus Memory geladen (kostet Latenz + Zeit für Datentransfer)
  - alle weiteren Zugriffe auf Cacheline sind schnell
  - Mittel, um das Latenzproblem zu mildern (s.u.)
- Cachelines werden nach einem bestimmten Algorithmus wieder aus dem Cache entfernt (zurückgeschrieben und/oder invalidiert)
- **Worst Case**: Laden einer Cacheline, von der nur ein einziges Element benutzt wird
  - tritt auf bei irregulärem Speicherzugriff (z.B. indirekt)

- **Caches sind schnell und klein, Speicher ist groß und langsam!**
- Folge: Langsame Datenpfade sind zu vermeiden!

```
DO I=1,N
  A(I)=B(I)+C(I)
ENDDO
DO I=1,N
  A(I)=A(I)*D(I)
ENDDO
```

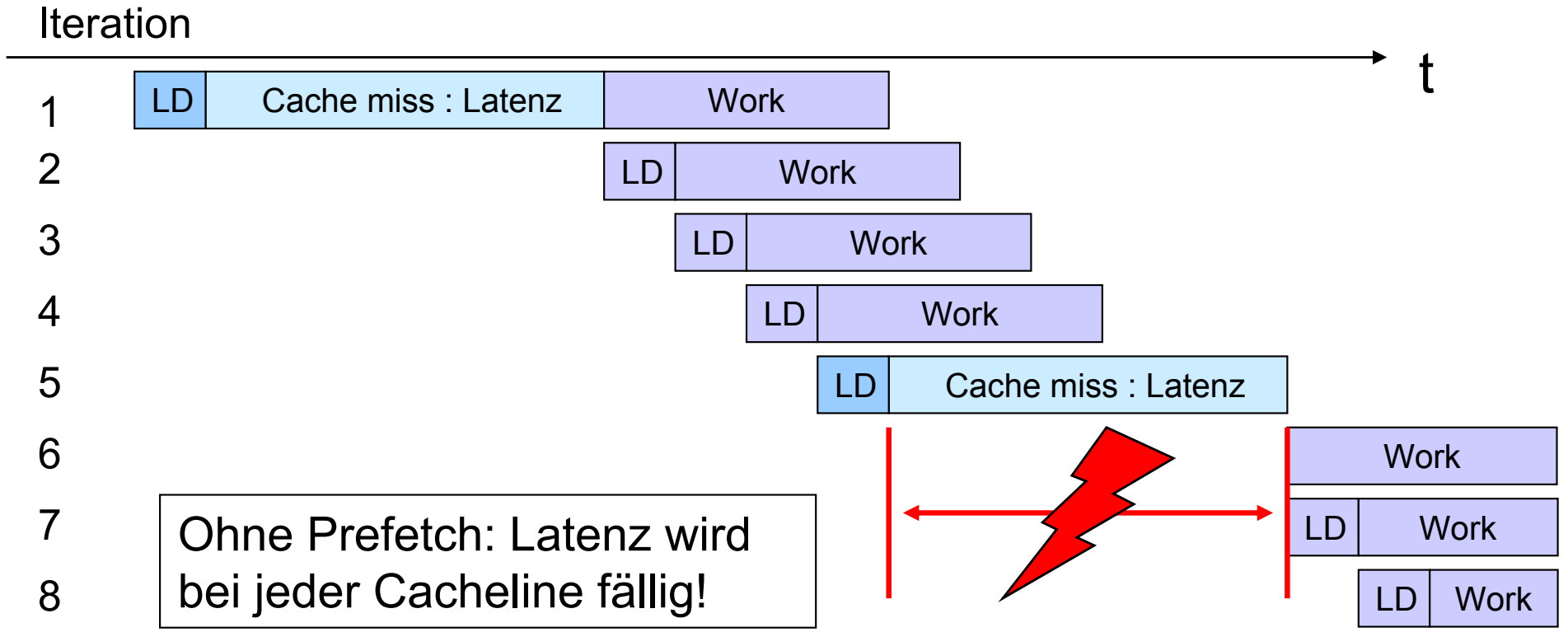
Optimierung

```
DO I=1,N
  A(I)=(B(I)+C(I))*D(I)
ENDDO
```

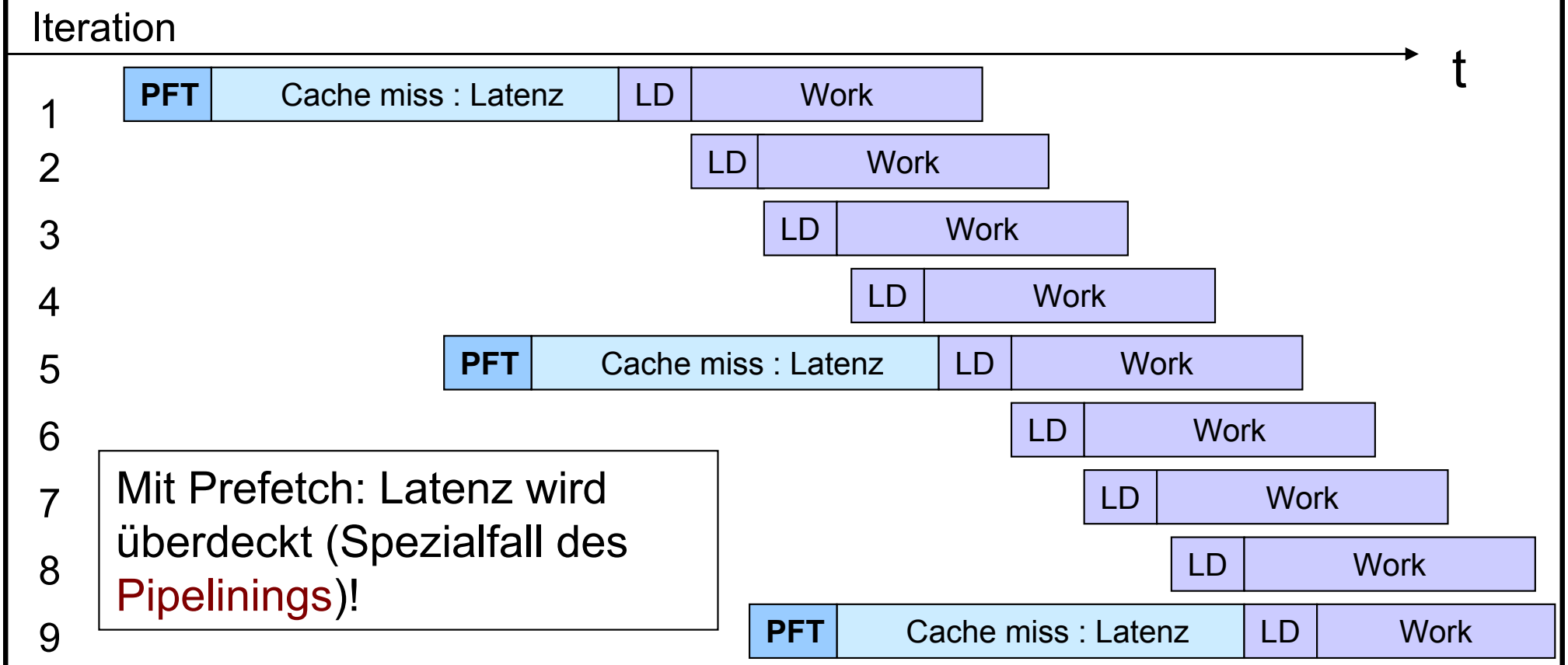
Effiziente Nutzung schneller Datenpfade ist absolut vorrangig!

Optimierung für Cache (2): Latenz überdecken mit Prefetching

- Wenn sich "Streaming" nicht vermeiden lässt (üblich in CFD), muss es wenigstens effizient geschehen: **Prefetching!**
  - **Prefetch** = Überdecken der Speicherlatenz durch rechtzeitiges Absetzen von Anfragen für Cachelines



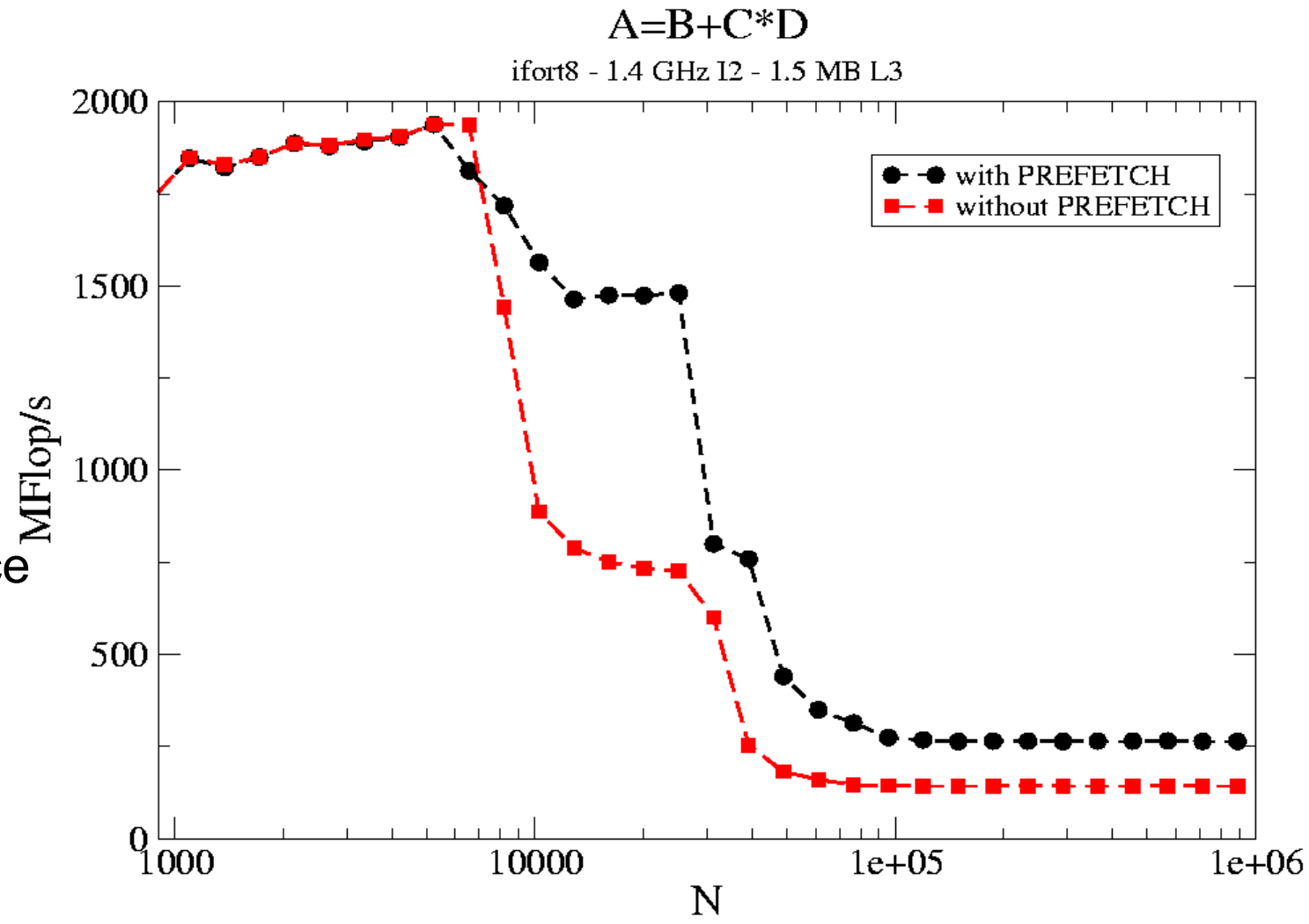
Optimierung für Cache (3): Latenz überdecken mit Prefetching



- Prefetching wird i.A. vom Compiler generiert (überprüfen!)
  - eingreifen u.U. im Quellcode möglich durch Compilerdirektiven
- Xeon, Opteron, Power4/5: zusätzlich **hardwarebasierter Prefetch**

Optimierung für Cache (4): Latenz überdecken mit Prefetching

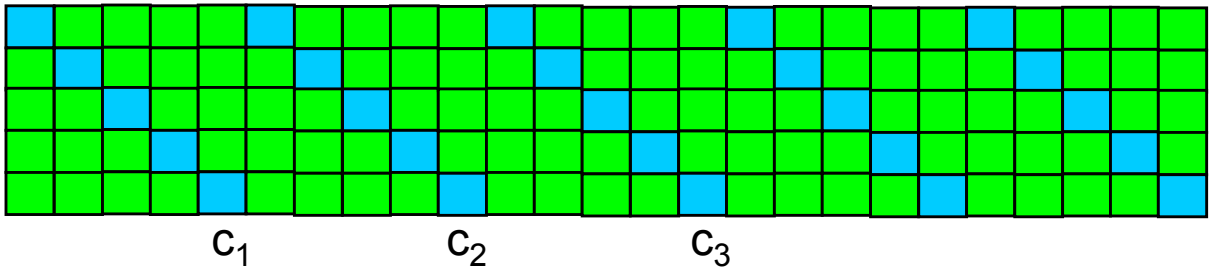
- Vektortriade als Beispiel
- Prefetch auch im Cache sinnvoll
- Verdoppelung der Performance im Speicher durch PFT



- Elementare Arithmetik: Pipelining ist das A und O
- Aufteilung einer komplexen Operation (hier Subtraktion) in:
  - compare exponent
  - shift mantissa
  - add mantissa
  - normalize exponent

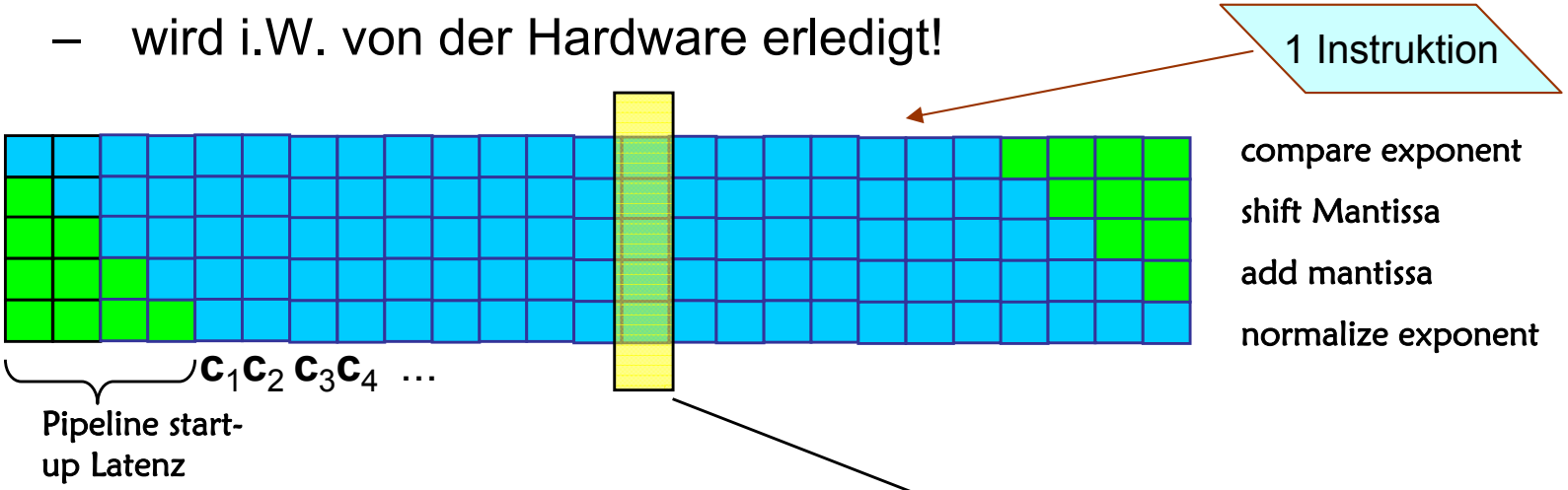
```
1.77e4 - 9.3e3
1.77e4 - 0.93e4
0.84e4
8.40e3
```

- Ablaufschema für Array-Operation (C = A + B)
  - hier: ohne Pipelining



compare exponent
shift Mantissa
add mantissa
normalize exponent

- Wie kann man das verbessern?
  - Pipelining: von 1 Ergebnis / 5 Takte zu einem Ergebnis / 1 Takt
  - wird i.W. von der Hardware erledigt!

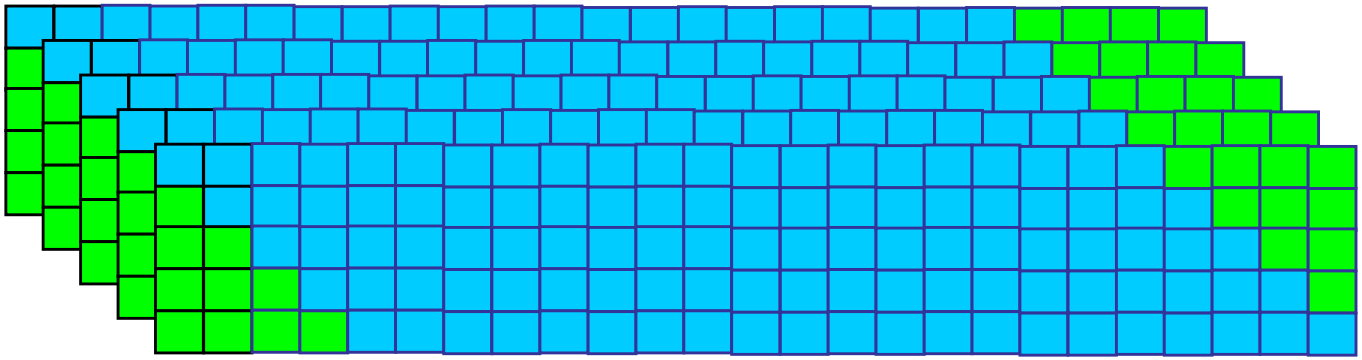


gleichzeitige Arbeit an Elementen
14, 13, 12, 11, 10

- Bei **RISC-CPUs** umso relevanter, je näher die Daten an den arithmetischen Einheiten liegen
- **Vektor-CPU**s haben die notwendige Speicherbandbreite, um Pipelines ohne Verzögerung zu "füttern"
  - Prefetch erfolgt dabei automatisch

Pipelining (3): Multi-Track Pipelines in Vektorrechnern

- Noch besser:
  - mehrere parallele Pipelines
  - $\mathbf{N_p}$ "Tracks" → $\mathbf{N_p}$ Ergebnisse pro Takt



- Voraussetzungen für gute Performance:
  - Pipes beschäftigt halten!
  - Compiler muss Vektorstrukturen im Code erkennen können
    - Unabhängigkeit aufeinander folgender Operationen

- Effizientes Pipelining kann durch Datenabhängigkeiten blockiert werden

Keine Abh.:

```
do i=1,N
   a(i) = a(i) * c
end do
```

Abhängigkeit:

```
do i=2,N
   a(i) = a(i-1) * c
end do
```

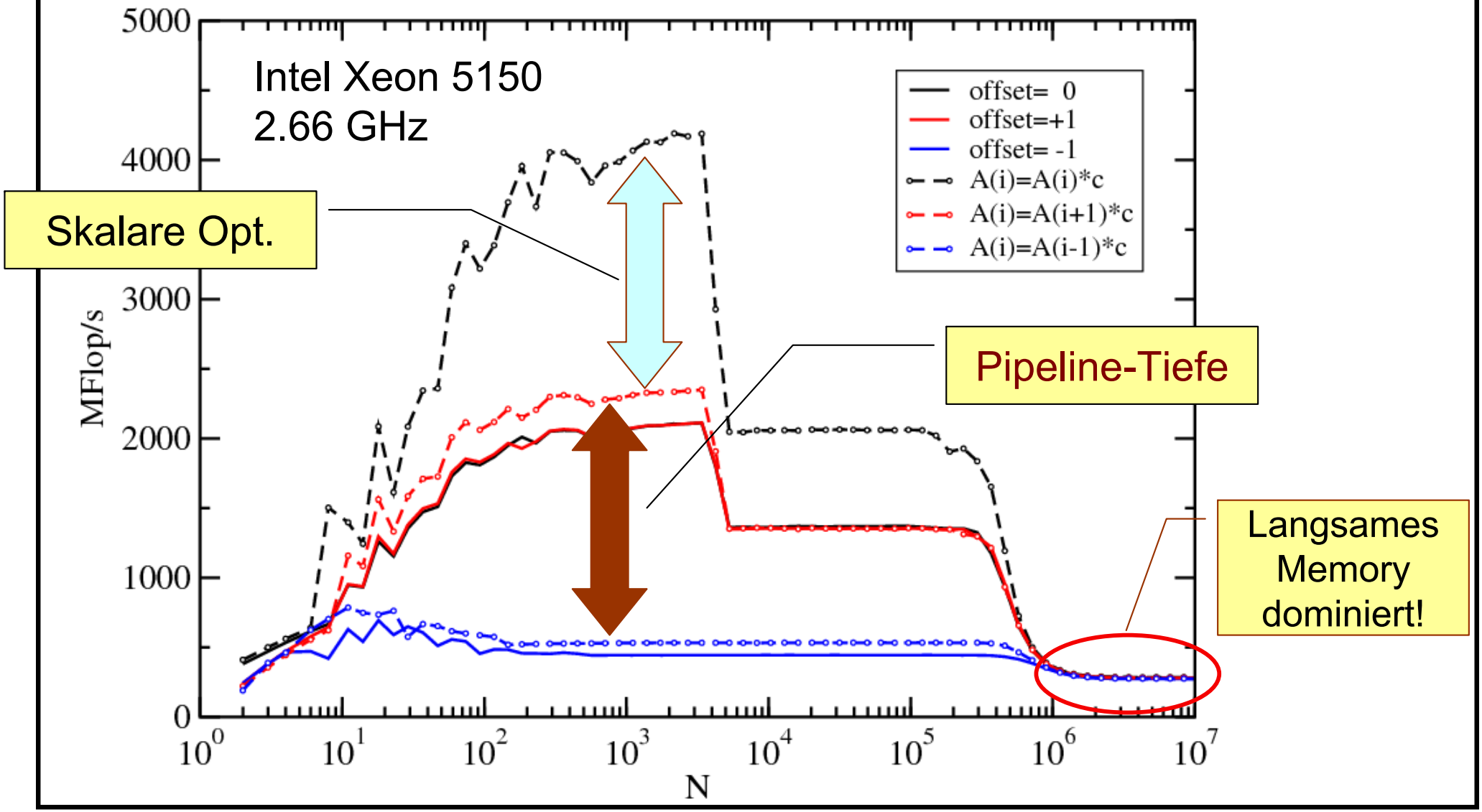
Pseudo-Abh.:

```
do i=1,N-1
   a(i) = a(i+1) * c
end do
```

Allgemeine Version (Offset als input-Parameter):

```
do i=max(1-offset,1),min(N-offset,N)
   a(i) = a(i-offset) * c
end do
```

Pipelining (5): Fallstricke



- Woher weiß man, ob ein Code die Ressourcen des Systems effizient nutzt?
- In vielen Fällen ist eine Abschätzung der zu erwartenden Performance aus den Eckdaten der Architektur und der Codestruktur möglich

- Randbedingungen der Architektur (maschinenabhängig):
  Speicherbandbreite GWorte/s
  Fließkommaleistung GFlop/s
  daraus abgeleitet:

  Maschinenbalance

$$B_m = \frac{\text{Speicherbandbreite}\,[Worte/s]}{\text{Fließkommaleistung}\,[Flop/s]}$$

- Typ. Werte (Hauptspeicher): **0.125 W/Flop (Itanium2 1.6 GHz)**
  0.06 W/Flop (Xeon 5160 3.0 GHz)
  0.5 W/Flop (NEC SX8)

- Zu erwartende Performance auf Schleifenebene?
- **Codebalance:**

$$B_c = \frac{\text{Datentransfer (LD/ST) } [Worte]}{\text{arithmetische Operationen } [Flops]}$$

- erwarteter Bruchteil der Peak Performance ("Lightspeed"):

$$l = \frac{B_m}{B_c}$$

- Beispiel?

Beispiel für Lightspeed-Berechnung: Vektortriade

- **Kernel benchmarks**:
  - Charakterisierung des Prozessors
  - Resultate können i.A. leicht verstanden und verglichen werden
  - Erlauben oft eine obere Schranke für die Leistungsfähigkeit eines Systems für eine spezifische Anwendergruppe abzuschätzen
  - Beispiele: *streams* , *cachebench*,…
- **Vektortriade**: Charakterisierung der erzielbaren Speicherbandbreiten

```
REAL*8 (SIZE): A,B,C,D
DO ITER=1,NITER
  DO i=1,N
     A(i) = B(i) + C(i) * D(i)
  ENDDO
  <OBSCURE>
ENDDO
```

Codebalance:

- Recheneinheiten:
  2 Flops / i-Iteration
- Bandbreite:
  (3 LD & 1 ST) / i-Iteration
- B_c **= 4/2 = 2 W/Flop**

Lightspeed-Berechnung: Vektortriade auf Itanium2 bei 1.6 GHz (1)

- **Funktionseinheiten: Performance allein durch die arithm. Einheiten beschränkt:** **2 MULTIPLY-ADD (2 MADD) Pipelines**

  Vektortriade:

  2 Iterationen ⟺ 1 Takt

  4 Flops / Takt ⟺ **Peak Performance**

  ➡ **Die Daten liegen jedoch i.A. nicht in den Funktionseinheiten/Registern, sondern im**

- **L2 Cache:** Bandbreite von

  2 LD und (2 LD oder 2 ST) pro Takt (=51.2 GB/s bei 1.6 GHz)

  Maschinenbalance: **4/4 = 1 W/Flop**

  Vektortriade:

  B_m/B_c = 0.5 ➡ **½ Peak Performance (3.2 GFlop/s)**

Lightspeed-Berechnung: Vektortriade auf Itanium2 bei 1.6 GHz (2)

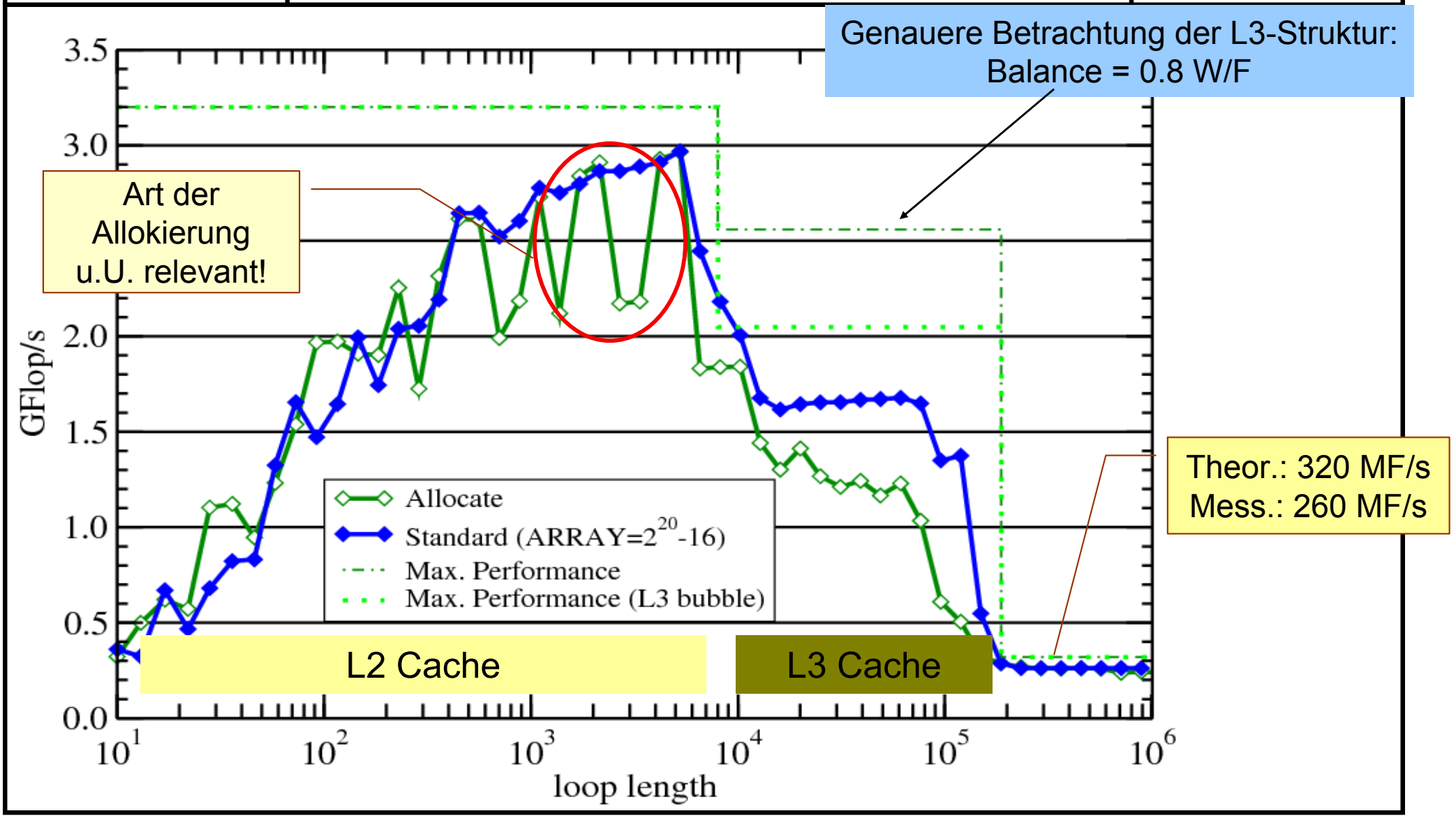
- **L3 Cache:** komplexe Situation, weil immer aus L2 geladen wird
  Abschätzung: halbe Performance wie aus L2

- **Speicher**: Bandbreite von
  - 2 LD oder 2 ST mit 400 MHz (Frequenz des Speicherbuses!)
    Maschinenbalance: 0.8/6.4 = 0.125 W/Flop
  - Achtung: "Read for Ownership" benötigt einen zusätzlichen LD vor jedem ST, d.h. **effektive Codebalance: B_c = 2.5 W/Flop**

  Vektortriade:

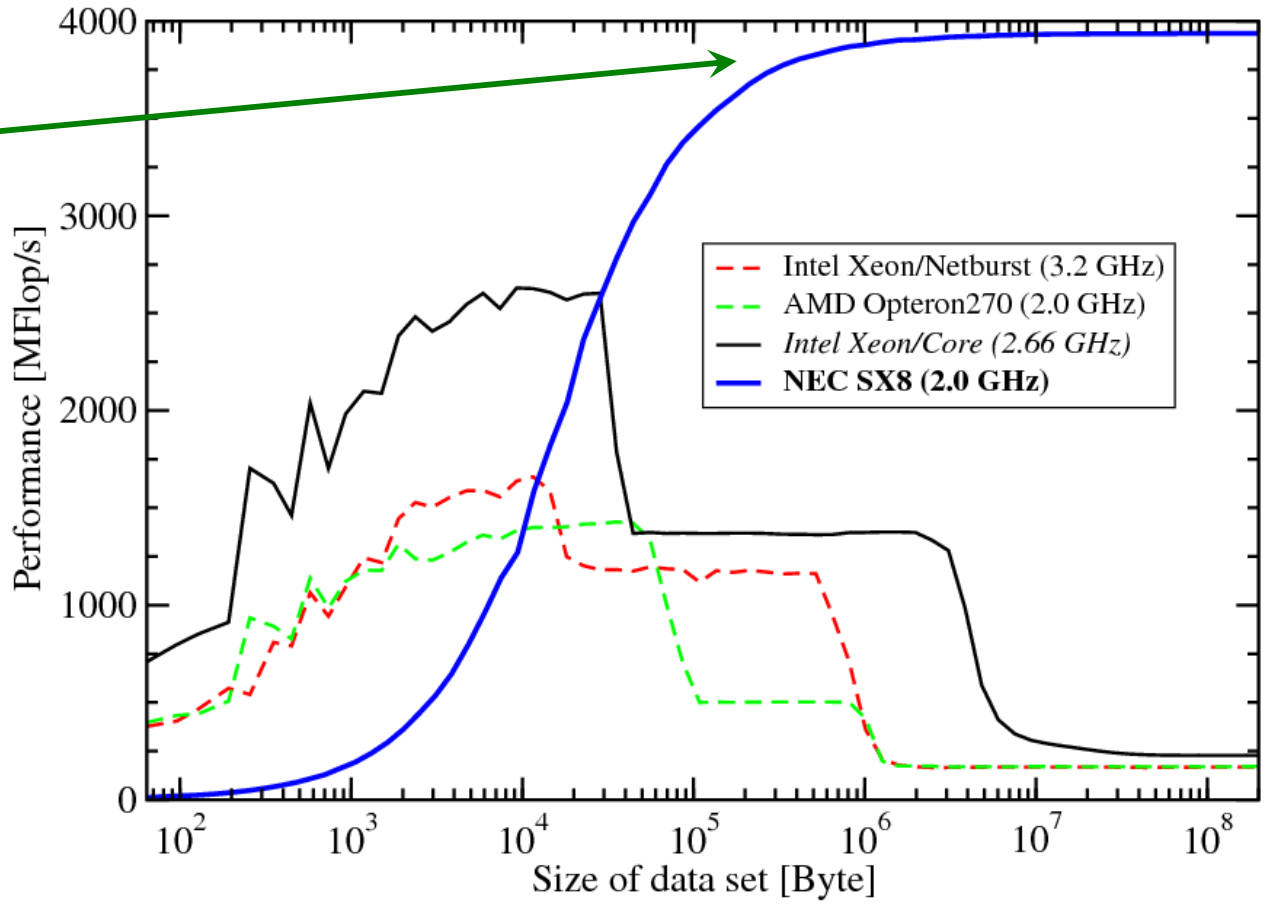
  B_m/B_c = 0.05 ⇒ **5% der Peak Performance (320 MFlop/s)**

Lightspeed-Berechnung: Vektortriade auf Itanium2 bei 1.6 GHz (2)



Kernel benchmarks: Vektor-Triade – Cache- vs. Vektorsysteme

NEC SX8
Max. Performance
4.0 GFlop/s



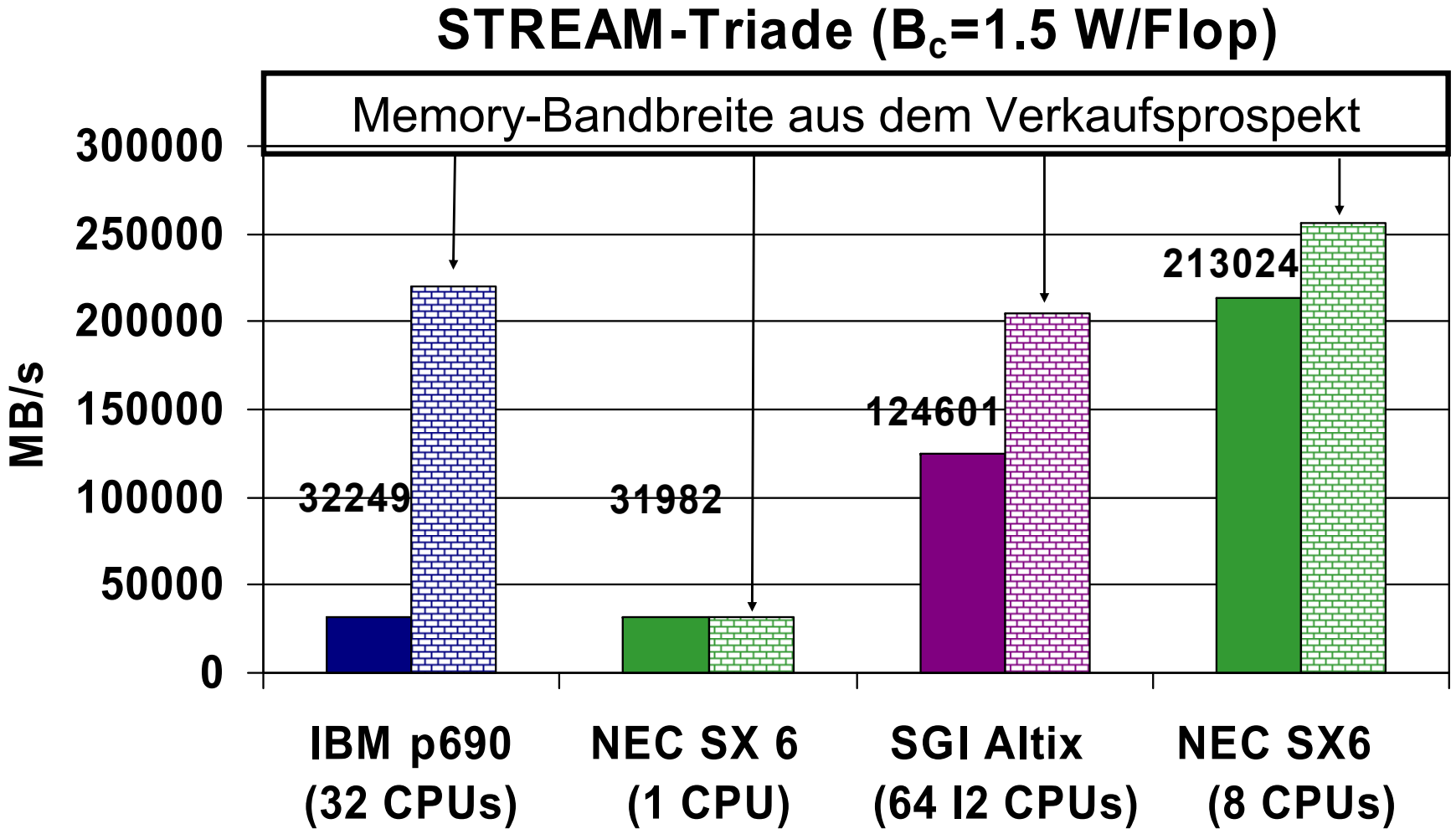
Welcher Rechner für welches Problem?

Cachebasierte Systeme:
Häufige Nutzung kleiner Datensätze

Vektorrechner:
U.U. unregelmäßiger Zugriff auf (sehr) große Datensätze

Optimierung in beiden Fällen notwendig!

Kernel benchmarks: *STREAM*
Dichtung und Wahrheit



http://www.cs.virginia.edu/stream/standard/Bandwidth.html

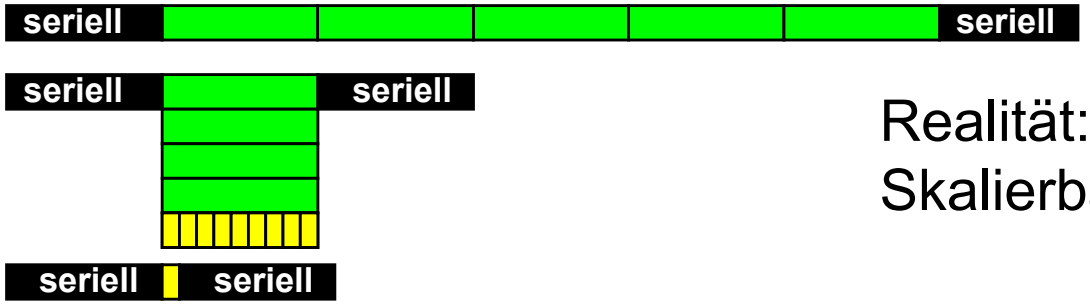
Möglichkeiten und Grenzen der Parallelität

"Parallelrechnen" = mehr als eine CPU zur Lösung eines numerischen Problems einsetzen

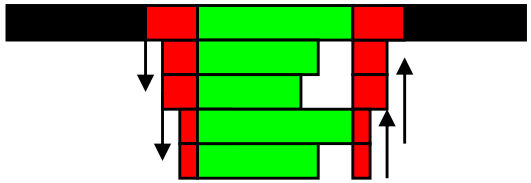
- Begriffe
  - **Skalierbarkeit** S(N): Wieviel mal schneller wird mein Problem gelöst, wenn es auf N CPUs läuft statt auf einer?
  - **Performance** P(N): Wieviele Rechenoperationen pro Sekunde macht das Programm mit N CPUs im Vergleich mit einer CPU?
  - **parallele Effizienz** ε(N)=S(N)/N: Welchen Bruchteil jeder CPU nutzt das parallele Programm?
- Viele numerische Aufgaben sind im Prinzip parallelisierbar
  - Lösen linearer Gleichungssysteme: **Ax = b**
  - Eigenwertberechnungen: **Ax = λx**
  - Zeitpropagation: **p(x,t+1)=$\mathcal{F}$(p(x,t))**

Grenzen der Parallelität

Idealfall: Arbeit ist komplett parallelisierbar



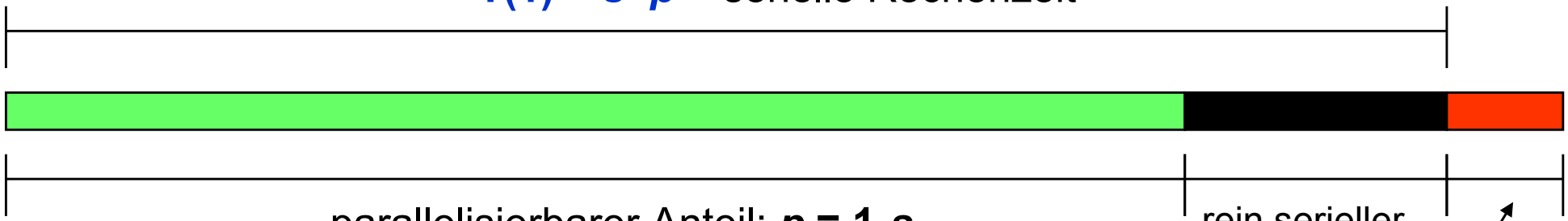
Realität: rein serielle Anteile begrenzen Skalierbarkeit



Kommunikation veschlimmert die Situation noch weiter

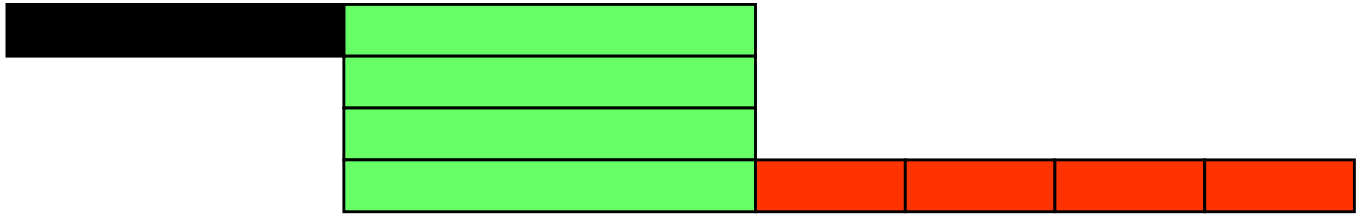
Grenzen der Parallelität: ein Modell

T(1) = *s*+*p* = serielle Rechenzeit



Bruchteil ***k*** für Kommunikation zwischen je 2 Prozessoren

parallel: T(*N*) = *s*+*p*/*N*+*Nk*



Allgemeine Formel für Skalierbarkeit (worst case):

Fall *k*=0: "Amdahl's Law" (strong scaling)

$$S_p^k(N) = \frac{T(1)}{T(N)} = \frac{1}{s + \frac{1-s}{N} + Nk}$$

- Limes großer Prozessorzahlen:
  - bei k=0: Amdahl's Law!

$$\lim_{N \to \infty} S_p^0(N) = \frac{1}{s}$$

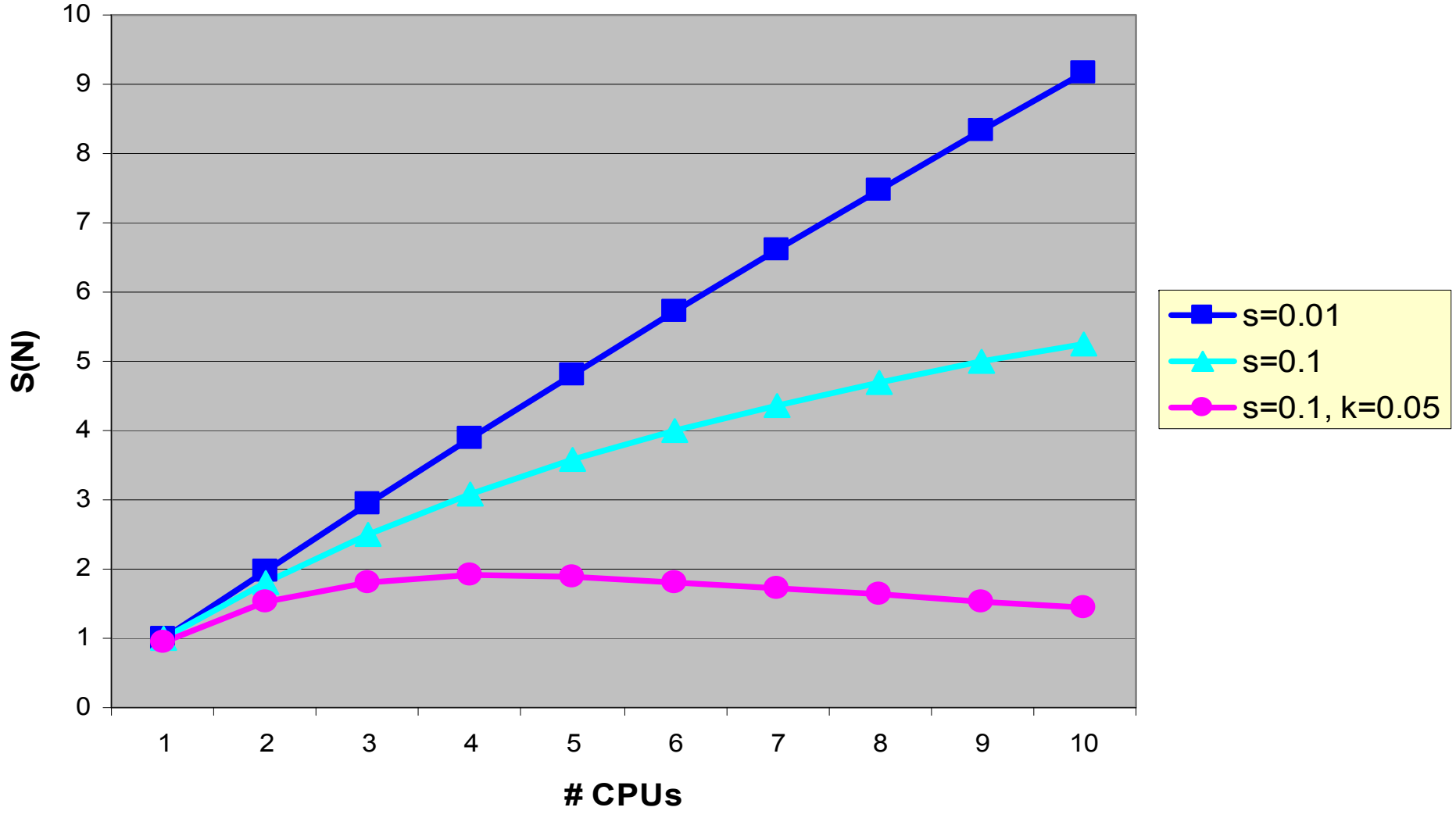
unabhängig von *N*

  - bei k≠0: einfaches Kommunikationsmodell liefert:

$$S_p^k(N) \xrightarrow{N>>1} \frac{1}{Nk}$$

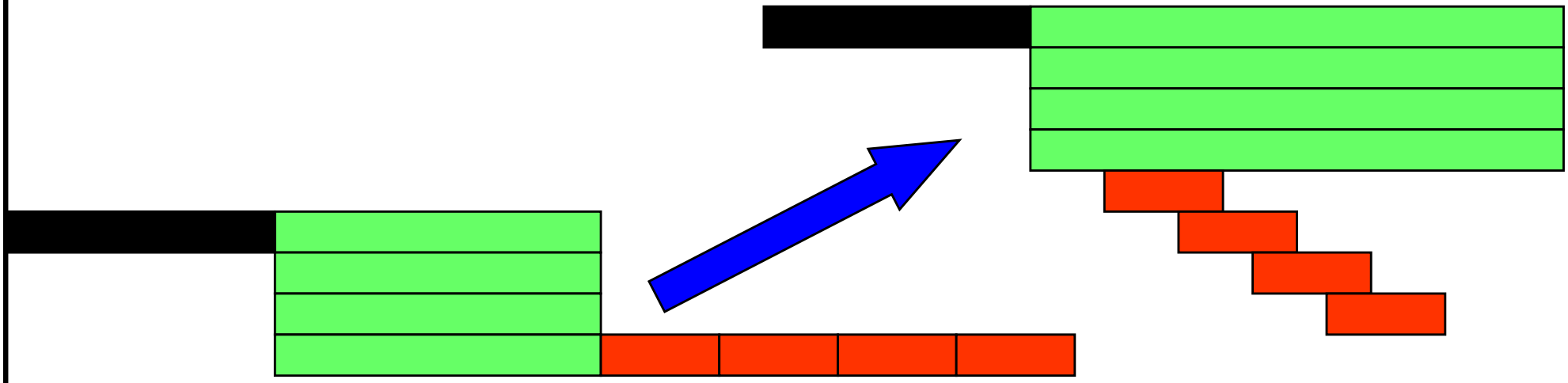
- Die Realität ist noch viel schimmer:
  - Load Imbalance
  - Overhead beim Starten paralleler Programmteile

Grenzen der Parallelität



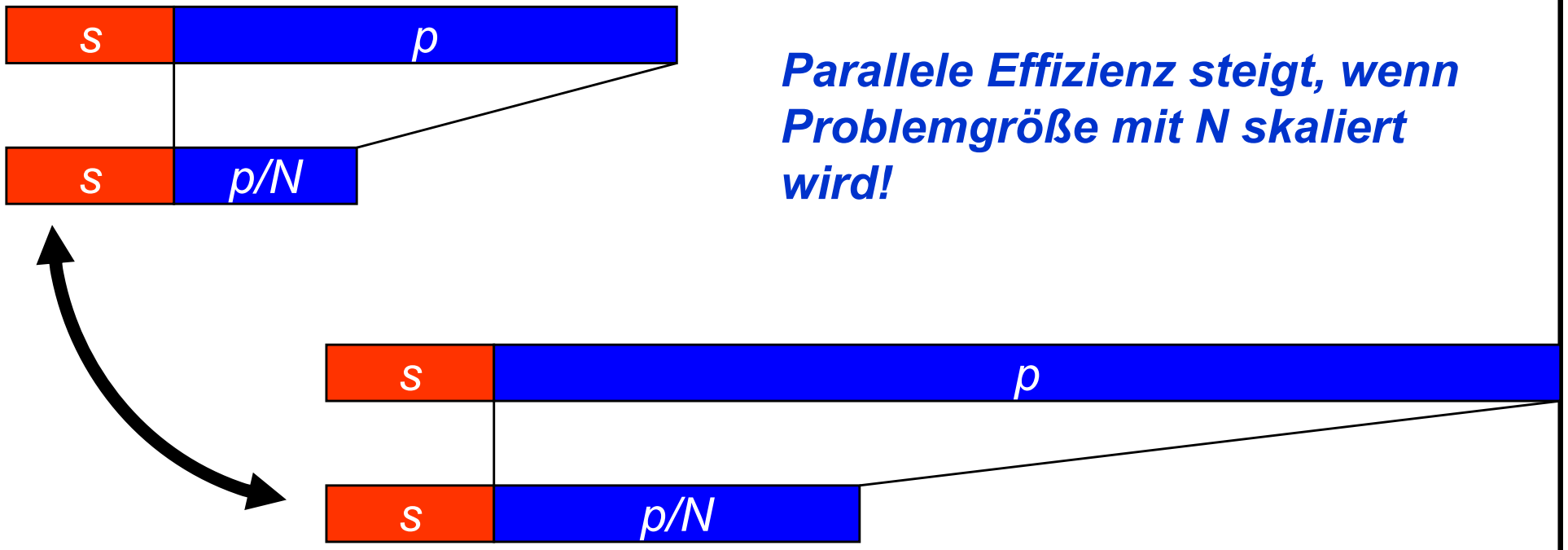
Grenzen der Parallelität: Ist Amdahl's Law unschlagbar?

- Kommunikation ist nicht immer rein seriell
  - nichtblockierende Netzwerke können gleichzeitig verschiedene Kommunikationsverbindungen schalten – Faktor *Nk* im Nenner wird zu log *k* oder *k* (technische Maßnahme)
  - u.U. kann Kommunikation mit nützlicher Arbeit überlappt werden (Implementierung, Algorithmus):



Grenzen der Parallelität: Erhöhen der parallelen Effizienz

- Vergrößern des Problems hat oft eine Vergrößerung von *p* zur Folge
  - *p* skaliert, während *s* konstant bleibt
  - Anteil von *s* an der Gesamtzeit sinkt
  - Nebeneffekt: Kommunikationsaufwand sinkt eventuell



Parallele Effizienz steigt, wenn Problemgröße mit N skaliert wird!

- Quantifizierung? Betrachte zunächst *k=0*!
  - Sei wie vorher ***s+p=1***
  - Perfekte Skalierung: Laufzeit bleibt konstant bei Vergrößerung von *N*
  - **Skalierbarkeit ist keine relevante Größe mehr!**
    „Parallele Performance“ =

$$\frac{\text{Arbeit/Zeit für Problemgröße } N \text{ mit } N \text{ CPUs}}{\text{Arbeit/Zeit für Problemgröße 1 mit 1 CPU}}$$

$$P_s(N) = \frac{s + pN}{s + p} = s + pN = s + (1 - s)N$$

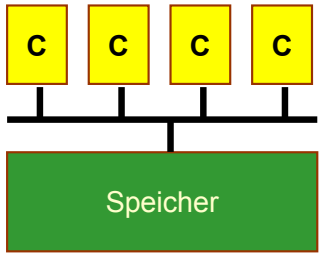
Gustafsson's Law ("weak scaling")

  - Linear in *N* – aber die Defnition der "Arbeit" spielt eine große Rolle!

Designprinzipien moderner Parallelrechner

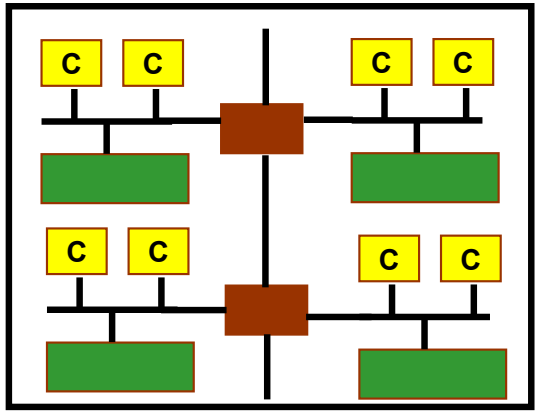
Moderne Parallelrechner Schematischer Aufbau

(1) Bus-basierte SMP-Systeme



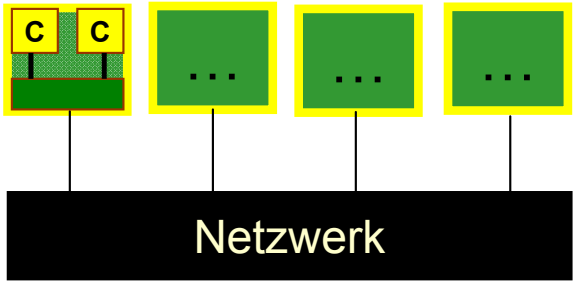
- 2-4 Prozessoren mit **einem** Speicherkanal
- Itanium / Xeon / Power4 / NEC SX

(2) Skalierbare „**Shared-memory**“ Systeme (**ccNUMA**)



- **Speicherbandbreite skaliert**
- Sehr schnelle interne Verbindungen (0.8 GB/s – 6.4 GB/s)
- SGI Origin / Altix, IBM p690 (ähnlich), Opteron-Knoten
- 4-256 Prozessoren

(3) Cluster von Rechenknoten (Architektur (1) oder (2))

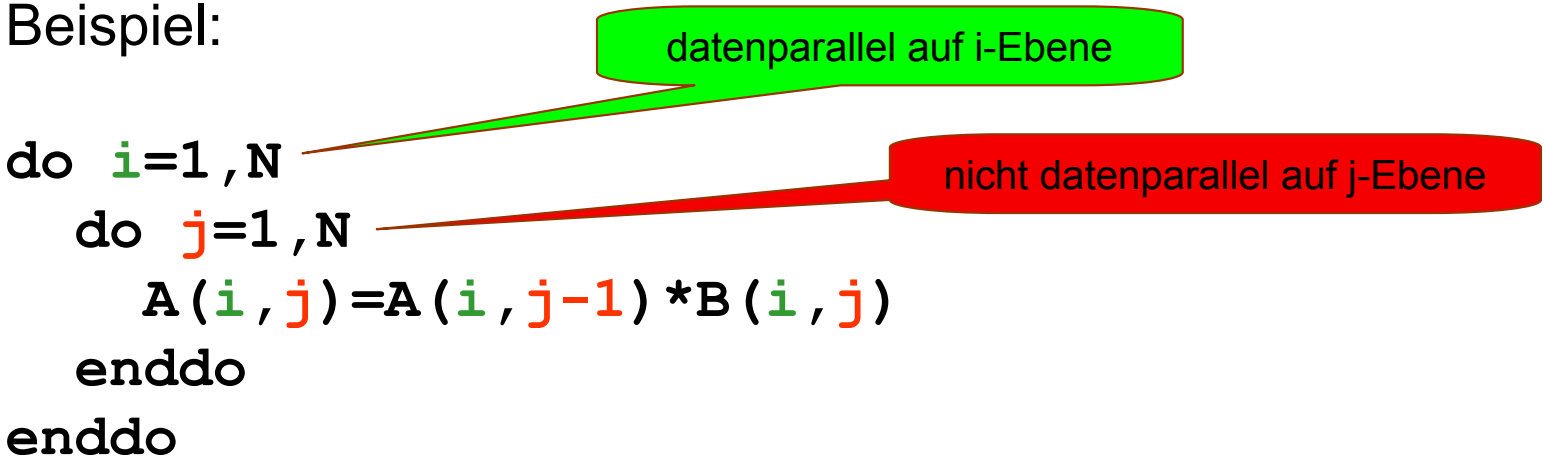


- **Speicherbandbreite skaliert, aber verteilter Speicher**
- Netzwerk: 50 – 1000 MByte/s
- 4 – 100000 Prozessoren

- Bei allen 3 Varianten (heute): **Multi-Core** möglich
  - d.h.: in einem "Sockel" steckt ein Chip mit 2 oder mehr (etwas langsameren) CPUs auf einem (oder 2) Stück Silizium
  - **Vorteil**: potenziell höhere Rechenleistung pro Sockel bei nahezu gleicher Leistungsaufnahme ("Macho-FLOPs")
  - **Nachteil**: Speicheranbindung der einzelnen CPU wird schlechter
- Dominant bei Parallelrechnern: **Cluster mit SMP-Knoten**
  - Kompromiss bzgl. Speicherbandbreite pro CPU
  - skalierend bis mehrere 100000 CPUs (akt. Rekord: **212992**)
- Spezialfall: "**Constellation**"-Systeme
  - CPUs/Knoten > # Knoten
  - **SGI Altix (HLRB II): ccNUMA-Architektur**

Parallele Programmierung in der Strömungsmechanik

- Parallele Programmierung beginnt immer mit der Identifikation von Parallelität im Code
  - **Datenparallelität**: Welche Teile eines großen Datensatzes (Feld, Matrix, …) können im Prinzip gleichzeitig bearbeitet werden? Beispiel:




```
do i=1,N
  do j=1,N
    A(i,j)=A(i,j-1)*B(i,j)
  enddo
enddo
```

  - **funktionelle Parallelität**: Welche funktionalen Einheiten im Programm können im Prinzip gleichzeitig ablaufen?
    - in der Praxis selten benutzt

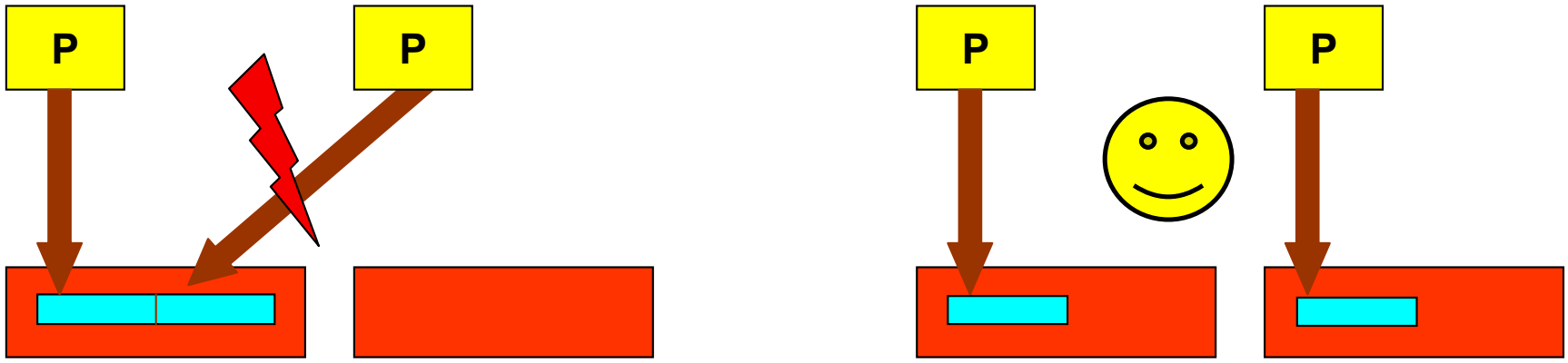
- Abhängig von der Rechnerarchitektur bieten sich 2 Zugänge an:
  - **Distributed-Memory-Systeme**

    Ein paralleles Programm besteht aus mehreren **Prozessen**. Jeder Prozess hat seinen eigenen Adressbereich und keine Zugriff auf die Daten anderer Prozesse. Notwendige Kommunikation findet durch Übermittlung von Messages statt. Standard: **MPI**

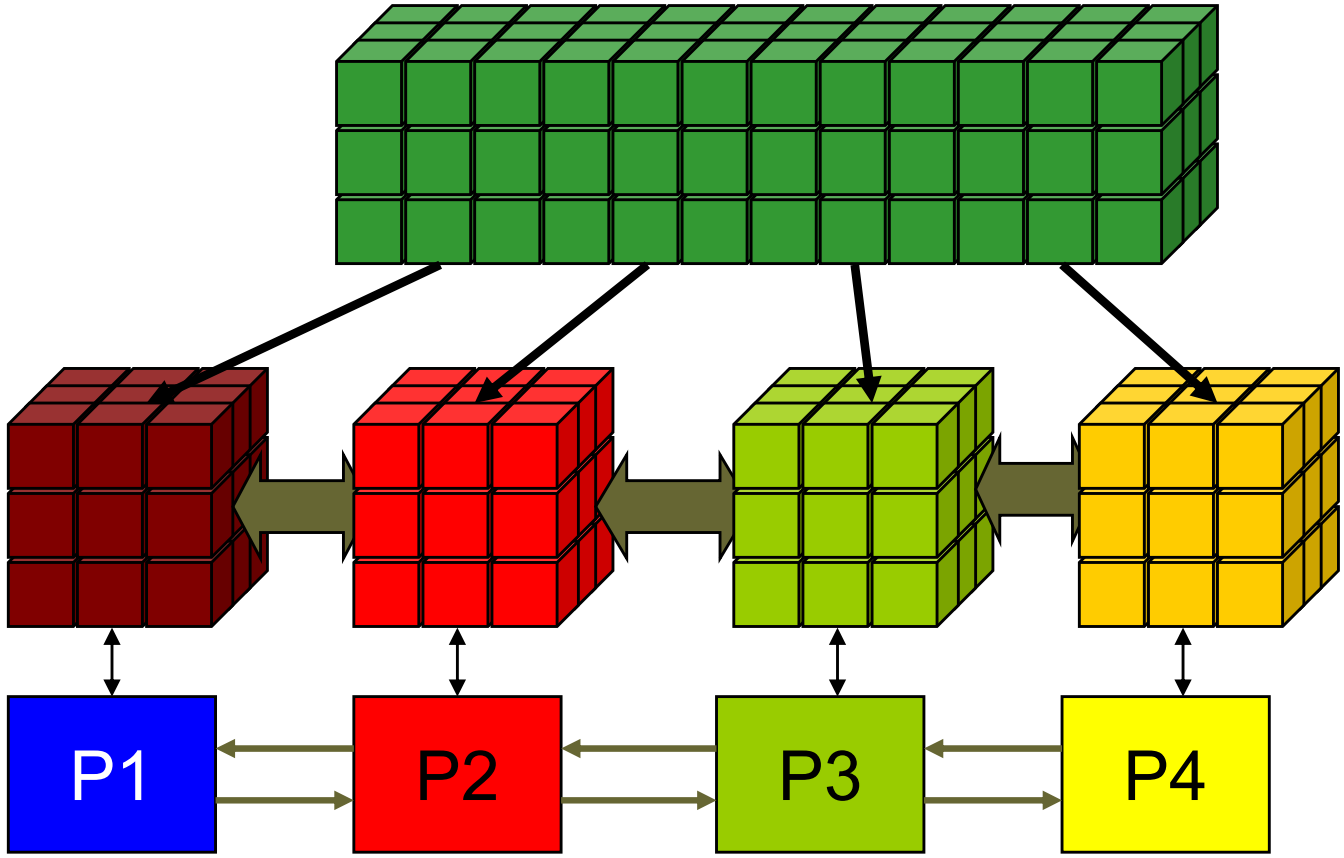
  - **Shared-Memory-Systeme (auch ccNUMA)**

    Ein paralleles Programm besteht aus mehreren **Threads**. Alle Threads haben Zugriff auf den kompletten Datenbereich. Standard: **OpenMP** (Compiler-Direktiven)

- Bei **ccNUMA-Systemen** (SGI Altix, Opteron) sind Verbindungen zwischen Knoten i.A. langsamer als die lokale Speicherbandbreite
- Folge: Es muss darauf geachtet werden, dass jede CPU auf den benötigten Speicher **möglichst lokal** zugreifen kann
  - Problem dabei: "**First Touch**" Policy mappt Speicherseiten dort, wo sie das erste Mal angefasst werden (i.A. bei der Initialisierung)
  - Wenn die Initialisierung nicht in gleicher Weise parallelisiert ist wie die Rechenschleifen, folgt eine sehr ungünstige Zugriffsstruktur:



- **Unabhängig von Programmiermodell**: Parallelisierung von CFD-Codes läuft oft über "**Domain Decomposition**"
  - Zerlegung des Rechengebietes in N Teile – eines für jede CPU

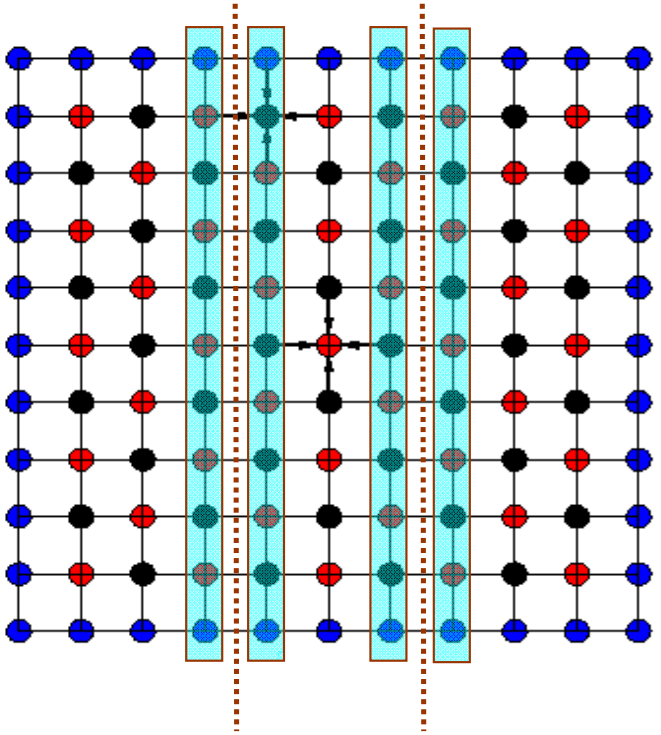


Beispiel für Gebietszerlegung: Red-Black Gauss-Seidel Solver

- Problem: Temperaturverteilung auf quadratischer Platte mit Randbedingungen
  - Lösung der Laplace-Gleichung:

$$\Delta T = 0$$

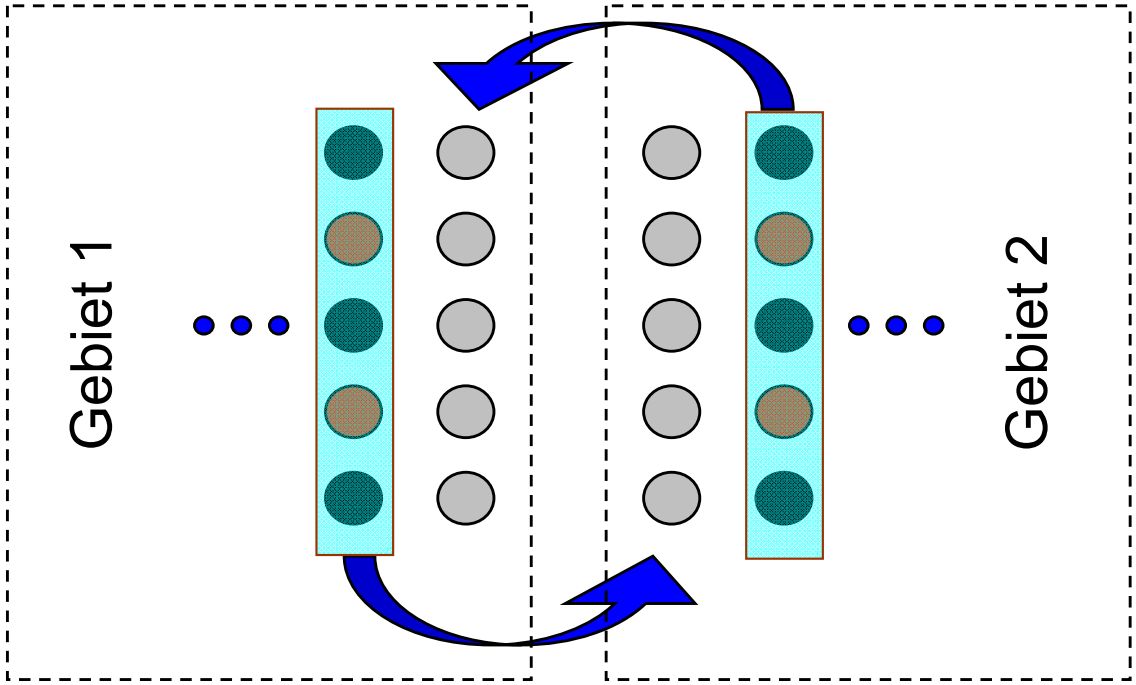
  - Numerik: Abb. auf NxN-Gitter
  - Algorithmus: "Relaxation"
    T an einem Punkt = Mittelwert der 4 Nachbarn
  - Iteration: Update erst der roten, dann der schwarzen Punkte



- Parallelisierung mittels Gebietszerlegung
  - Was macht man mit den Randzellen eines Gebietes?

Beispiel für Gebietszerlegung : Red-Black Gauss-Seidel Solver

- Lösung des Randzellenproblems durch "Geisterzellen" (**ghost cells**)
  - jedes Gebiet hat eine zusätzliche Schicht Randzellen, wo es an ein anderes Gebiet grenzt
  - Nach jeder Iteration (red/black Sweep) werden die Geisterzellen mit dem Inhalt der Randzellen des Nachbarn **beschrieben**



- Wirkt sich die Kommunikation auf die Skalierbarkeit aus?
  - **Annahmen**: keine Überlappung zwischen Kommunikation und Rechnung, nichtblockierendes Netzwerk, 3D-Problem
- 2 Fälle:

(1) **strong scaling**: Problemgröße konstant, N steigt → Latenz *L* wird wichtig

Kommunikationszeit = ***k/N$^{2/3}$+L***
Skalierung sättigt langsamer und auf niedrigerem Niveau!

$$S_p(N) = \frac{1}{s + \frac{1-s}{N} + kN^{-2/3} + L} \xrightarrow{N>>1} \frac{1}{s+L}$$

(2) **weak scaling**: Problemgröße steigt proportional zu N

Arbeit pro CPU bleibt konstant, ebenso die Kommunikation ***k***
Performance skaliert langsamer, aber noch linear

$$P(N) = \frac{s + pN}{s + p + k} = \frac{s + (1-s)N}{1+k}$$

CFD Applikationsperformance

SIP Solver nach Stone

- *Strongly Implicit Procedure* (SIP) nach Stone wird in Finite-Volumen Paketen oftmals zur Lösung des linearen Gleichungssystems

  A x = b verwendet

- Beispiele:
  - **LESOCC, FASTEST, FLOWSI** (LSTM, Erlangen)
  - **STHAMAS3D** (Kristalllabor, Erlangen)
  - **CADiP** (Theoret. Thermodynamik und Transportprozesse, Bayreuth)
  - **…**
- SIPSolver: **1) Unvollständige LU-Zerlegung von A**

  2) Serie von Vorwärts-/Rückwärts-Substitutionen und Residuenberechnungen
- Ursprüngliches Testprogramm (M. Peric):
  `ftp.springer.de:/pub/technik/peric`
- Optimierung: **HPC-Gruppe RRZE**

Vorwärts-Substitution (naive 3D-Version)

Datenabhängigkeit: `(i,j,k)` ⟵ `{(i-1,j,k);(i,j-1,k);(i,j,k-1)}`

```
do k = 2 , kMax
   do j = 2 , jMax
      do i = 2 , iMax
               RES(i,j,k) ={RES(i,j,k)
$              - LB(i,j,k)*RES(i,j,k-1)
$              - LW(i,j,k)*RES(i-1,j,k)
$              - LS(i,j,k)*RES(i,j-1,k)
$                          }*LP(i,j,k)
      enddo
   enddo
enddo
```

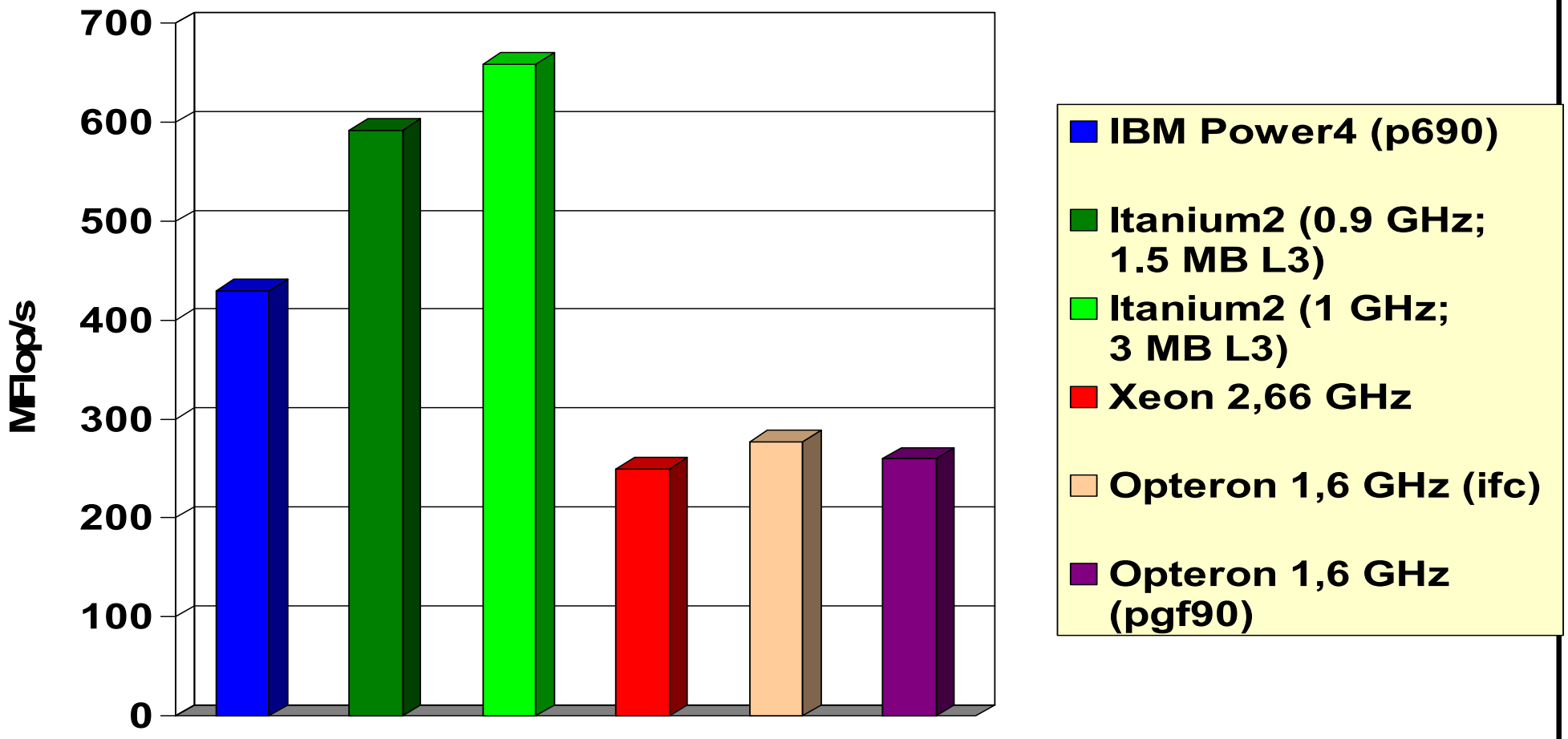
Datenabhängigkeit

- verhindert Vektorisierung oder einfache Parallelisierung
- erlaubt hohe örtliche Datenlokalität: Alle Elemente der Cachelines werden genutzt.
- erlaubt re-use von Cachelines: `RES(i,j,k),RES(i-1,j,k)`

➡ Geeignete Implementierung für **Cache-basierte Prozessoren!**

SIPSolver: Naive 3D-Implementierung: Einzelprozessorperfromance

- Gitter=91^3 (Problemgröße: 100 MB); 3D Version & naive Compiler-switches
- Performance: 1 Flop/Wort ➡ Obere Grenze für Itanium2: ~800 MFlop/s



SIPsolver: *hyperplane* Implementierung

Elimination der Datenabhängigkeit `(i,j,k)<-{(i-1,j,k);..}` durch Umgruppierung:

Innerste Schleife über eine *hyperplane*:

Alle Punkte `i+j+k=l`(mit festem `l`)

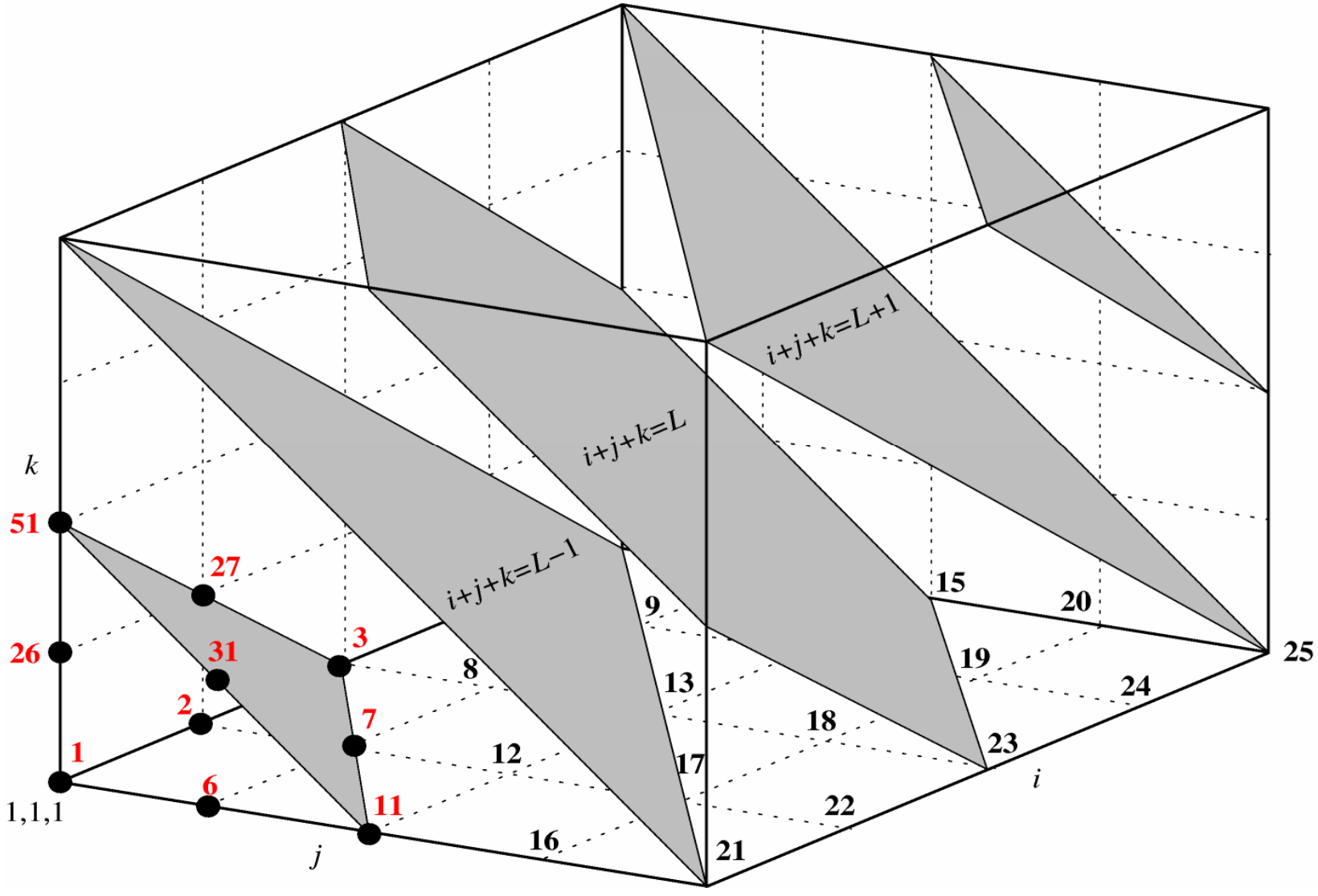
- liegen in der *hyperplane* `l`,
- sind voneinander unabhängig
- greifen nur auf Datenpunkte in der *hyperplane* `l-1` (Vorwärts-substitution) zu.

```
do l=1,hyperplanes
      n=ICL(l)
      do m=n+1,n+LM(l)
       ijk=IJKV(m)
       RES(ijk) =(RES(ijk) -
$      LB(ijk)*RES(ijk-ijMax)-
$      LW(ijk)*RES(ijk-1) -
$      LS(ijk)*RES(ijk-iMax) )
$      *LP(ijk)
      enddo
enddo
```

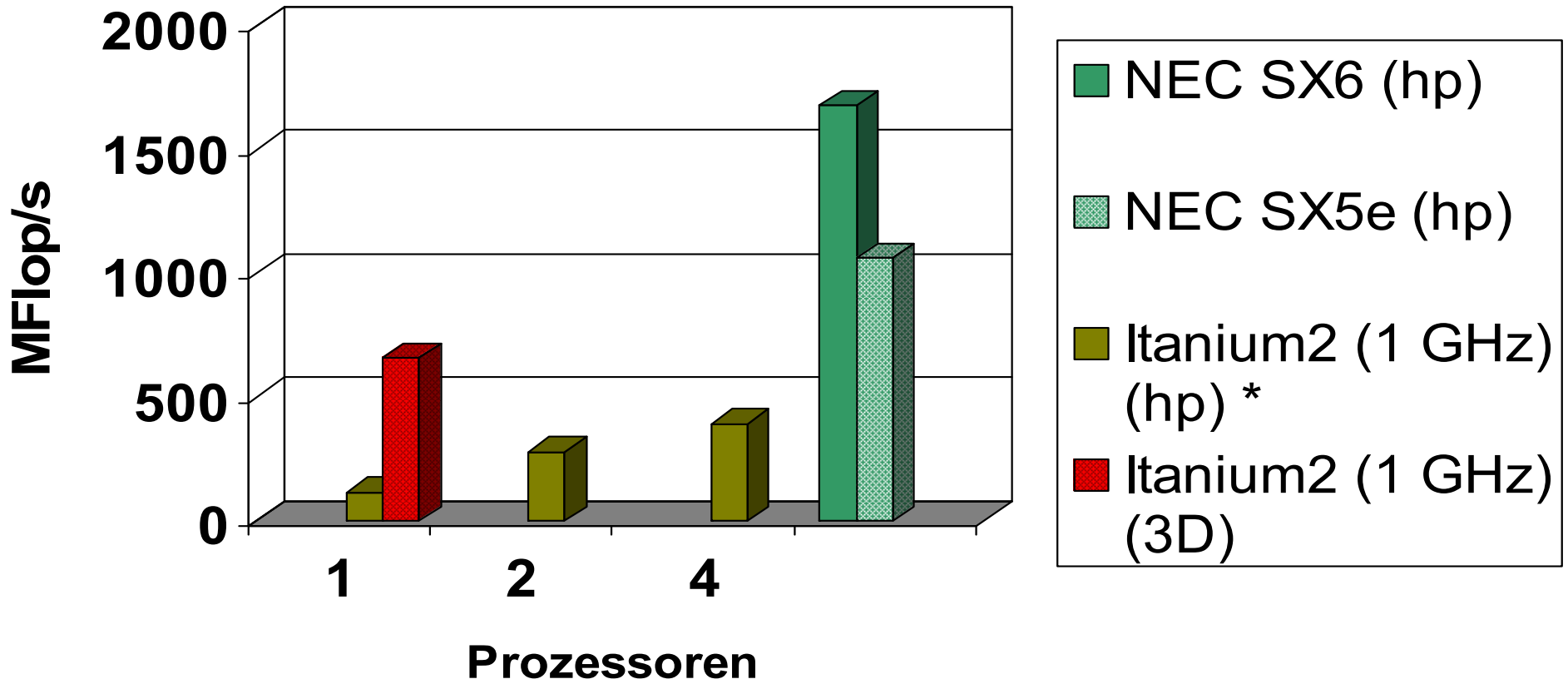
- Einfache Vektorisierung/ Parallelis. der inneren Schleife
- Verlust der Datenlokalität

→ Geeignete Implementierung für **Vektorprozessoren**

SIPsolver:
hyperplane Implementierung



SIPSolver: *hyperplane* Leistungszahlen



- Gute Skalierung für Itanium2 SMP System, aber schlechte absolute Performance
- Gute absolute Einzelprozessorperformance für NEC SX6 Vektorprozessor

SIPSolver: *pipeline parallel processing* Idee und Implementierung

Ziel:

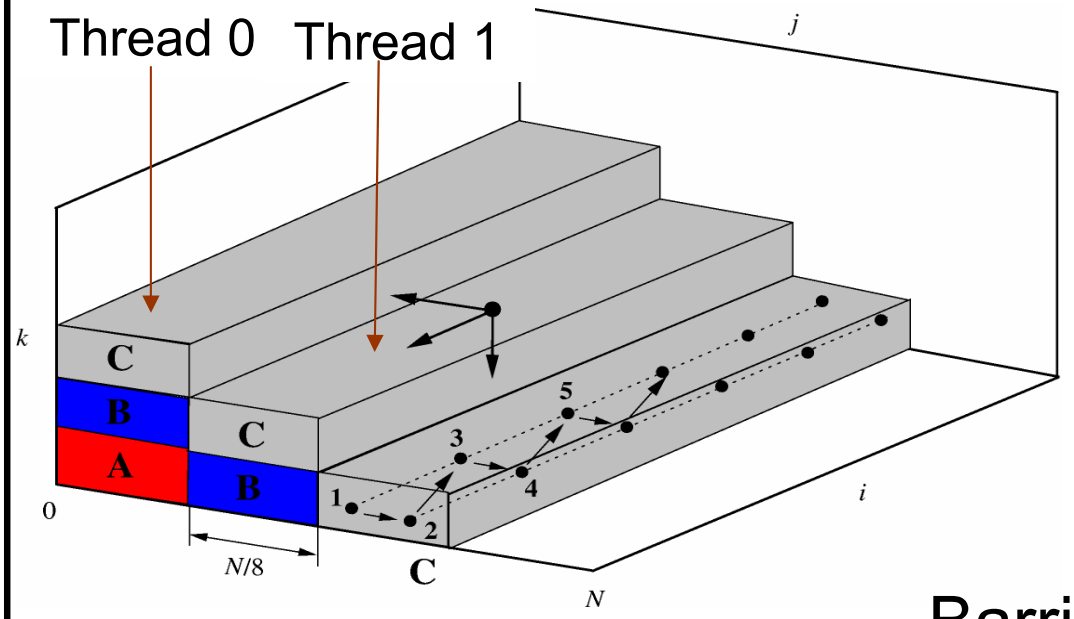
Parallelisierung (Shared-Memory) und Erhalt der Datenlokalität der 3D-Implementierung

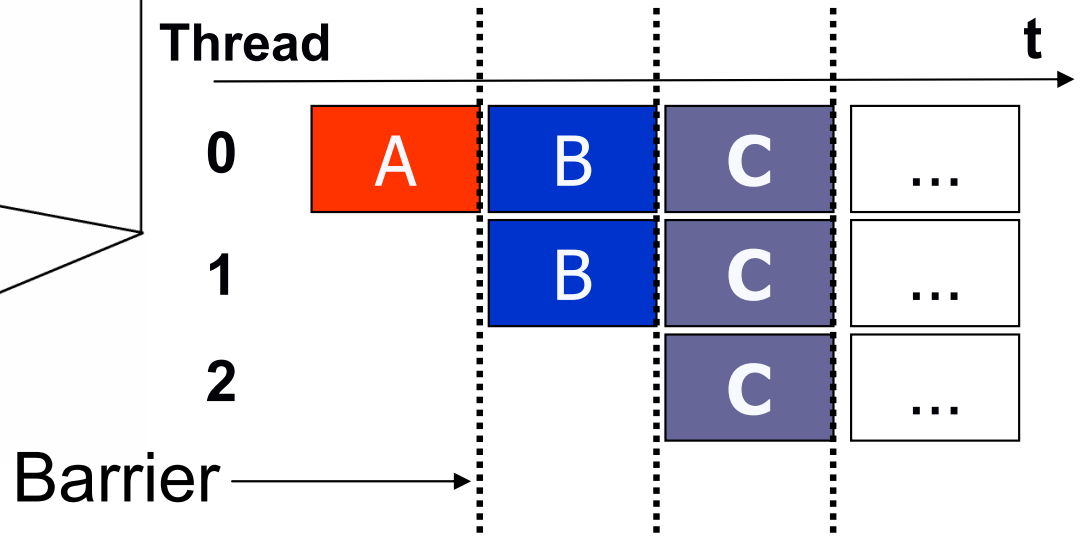
Vorgehensweise:

- 3D-Implementierung `(i,j,k)<-{(i-1,j,k);..}`
- Parallelisierung in `j` Richtung
- Aufheben der Datenabhängigkeiten durch Synchronisieren nach jedem Schritt in `k` - Richtung

```
$omp parallel private(….)
do l =2, kMax+numThreads-2
   threadID=OMP_GET_THREAD_NUM()
   k = l - threadID
   if((k.ge.2).and.(k.le.kMaxM)) then
   do j = jS(threadID),jE(threadID)
      do i = 2 , iMax
        RES(i,j,k)={RES(i,j,k-1)-
$       LB(i,j,k)*RES(i,j,k-1)
$       …
      enddo
   enddo
   endif
$omp barrier
enddo
$omp end parallel
```

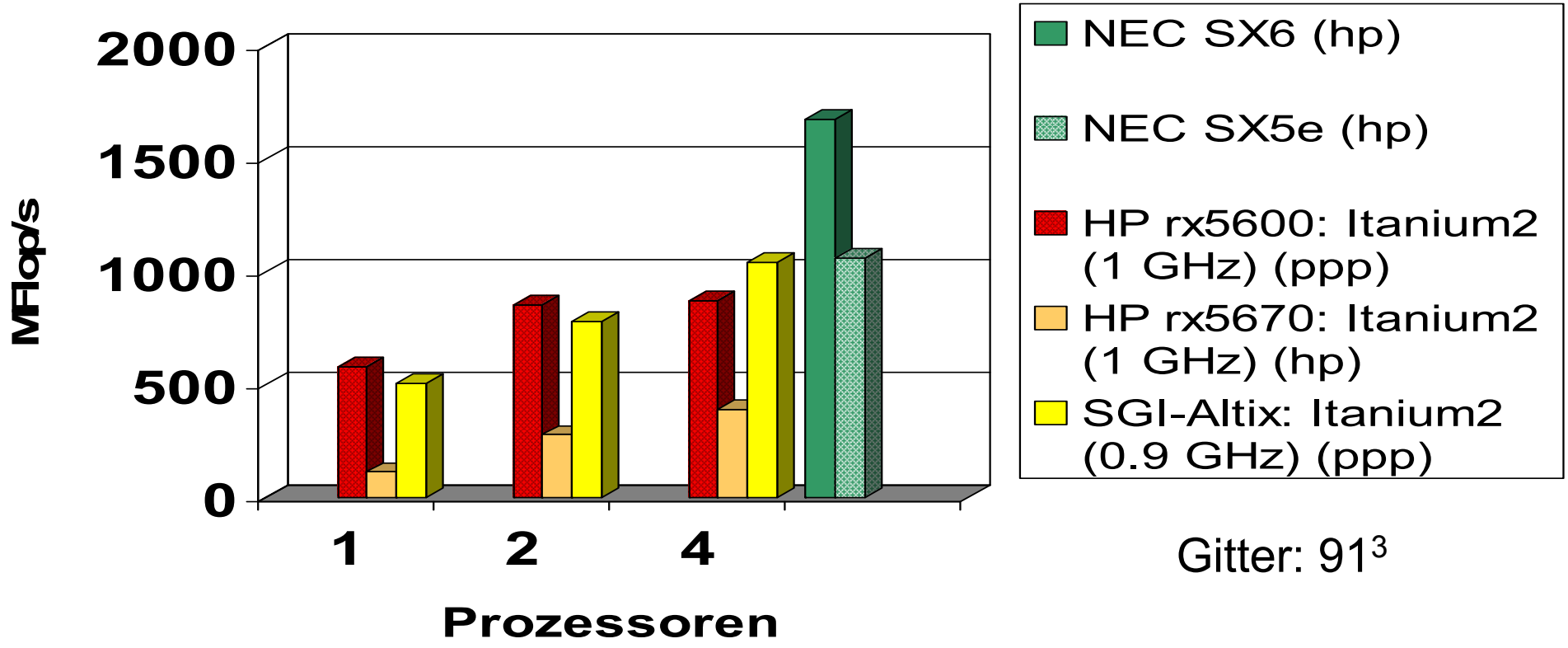
SIPSolver: *pipeline parallel processing* Schematische Darstellung





- Parallelisierung sinnvoll, falls **`kMax, jMax >> #Threads`**
- Automatische Parallelisierung der 3D Version durch Hitachi-Compiler!
- Beste Version für große SMP Knoten mit Cache-basierten Prozessoren

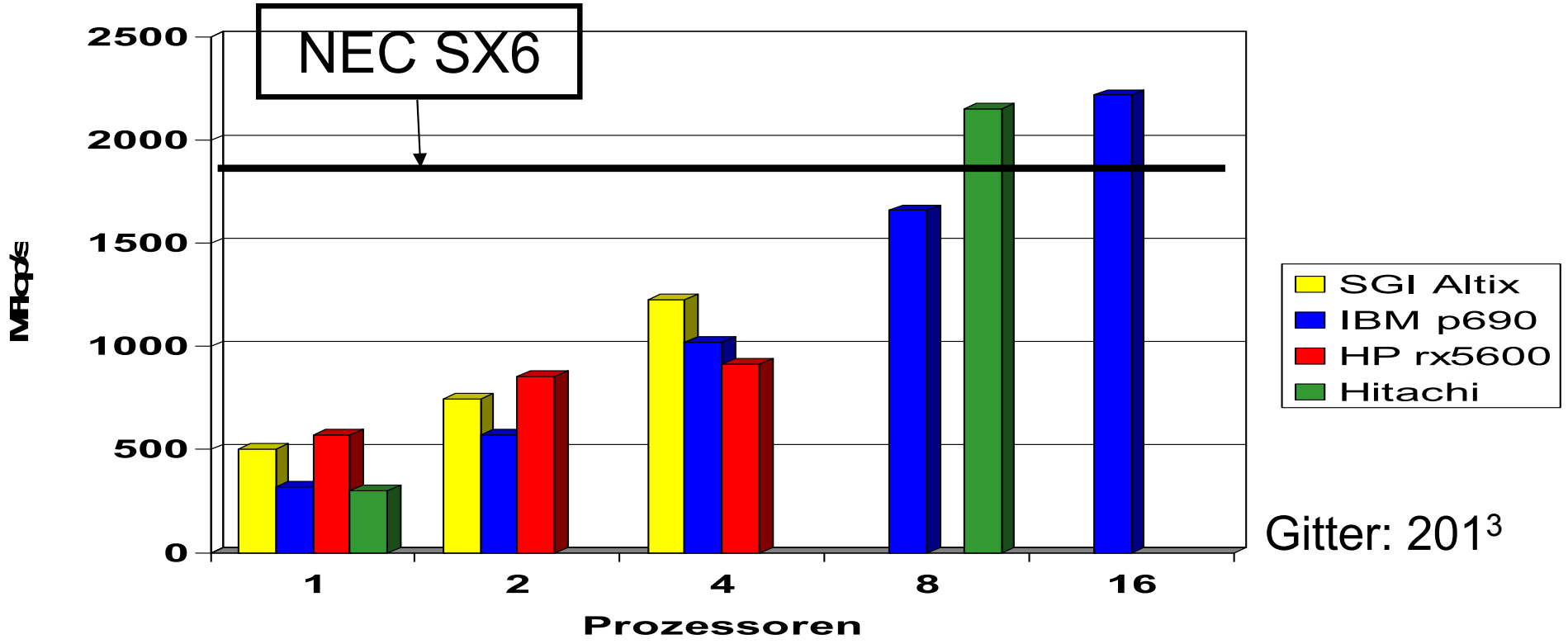
SIPSolver: SMP-Skalierbarkeit (Pipeline Parallel Processing)



Gitter: 91^3

- HP rx5670: 4-fach Itanium2-System mit nur einem Speicherkanal für 4 Prozessoren (Speed-Up(4)=1,5)
- SGI Altix: 32-fach Itanium2-System mit einem Speicherkanal für **jeweils** 2 Prozessoren (Speed-Up(4)=2,0)

SIPSolver:
SMP Performance vs. Vektorprozessor



Gitter: 201^3

- IBM p690: 32-fach Power4-System mit einem Speicherkanal für **jeweils** 2 Prozessoren (Speed-Up(16)=7)
- Hitachi SR8000: 8-fach „Power3“-System mit einem Speicherkanal pro Prozessor (Speed-Up(8)=7.2)

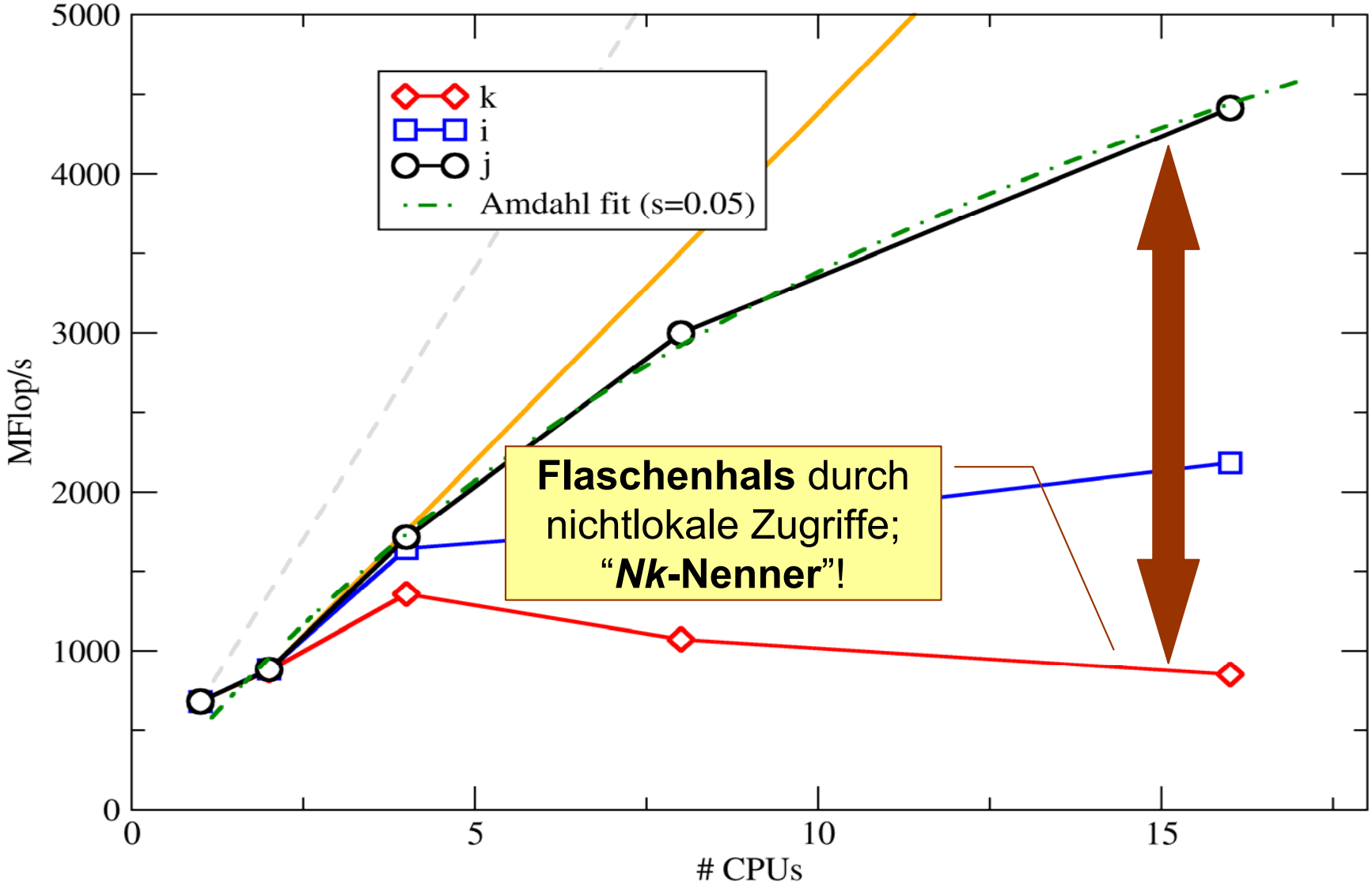
- **Auf ccNUMA-Systemen**: korrekte Parallelisierung der Initialisierungsschleife ist extrem wichtig!

```
!$omp parallel do
      do k=1,kMax
         do j=1,jMax
            do i=1,iMax
               T(i,j,k)=0.
            enddo
         enddo
      enddo
```

```
                  do k=1,kMax
                     do j=1,jMax
               !$omp parallel do
                           do i=1,iMax
                              T(i,j,k)=0.
                           enddo
                     enddo
                  enddo
```

```
                  do k=1,kMax
            !$omp parallel do
                     do j=1,jMax
                        do i=1,iMax
                           T(i,j,k)=0.
                        enddo
                     enddo
                  enddo
```

SIP-solver (ppp): NUMA Placement
Einfluss der Init-Parallelisierung auf Altix



Ergebnisse für CFD-Applikationen:

- Speicherbandbreite ist die entscheidende Einflussgröße
- Effiziente Nutzung schneller, Vermeidung langsamer Datenpfade ist Pflicht!
- Pipelining im Speicherzugriff und den arithmetischen Einheiten ermöglichen!
- Parallelisierung muss immer auch parallele Effizienz berücksichtigen
  - Kommunikationsaufwand beachten
  - strong/weak scaling
- Parallele Skalierbarkeit der Problemstellung ist Voraussetzung, um SMP Systeme/Cluster in Konkurrenz zu Vektorprozessoren zu setzen
- Bei ccNUMA-Systemen ist auf korrektes Placement zu achten!

Vielen Dank!

HPC-Dienste und Rechner:

- RRZE: **http://www.hpc.rrze.uni-erlangen.de/**
- LRZ: **http://www.lrz-muenchen.de/services/compute/hlr/**
- HLRS: **http://www.hlrs.de/**
- JSC: **http://www.fz-juelich.de/jsc/**

Dokumentation und Optimierungs-Manuals bei Prozessorfirmen:

- **http://developer.intel.com/products/processor/manuals**
- **http://developer.amd.com/documentation/guides**

HPC allgemein:

- KONWIHR: **http://konwihr.in.tum.de/**
- TOP500: **http://www.top500.org/**

Literatur:

- W. Schönauer: *Scientific Supercomputing*. http://www.rz.uni-karlsruhe.de/~rx03/book/
- Dowd/Severance: *High Performance Computing*. 2nd edition, O'Reilly, 1998
- Hager/Wellein: *Concepts of High Performance Computing*. RRZE, 2007 (a.A.)
- **http://www.blogs.uni-erlangen.de/hager/topics/Publications/**