Die Marvel, ein gedrosselter Supercomputer

- Warum ist die Marvel so schnell?
- Warum ist die Marvel so langsam?
- Erfahrungen mit dem Softwaresupport

Warum ist die Marvel so schnell?
-- Hardware --

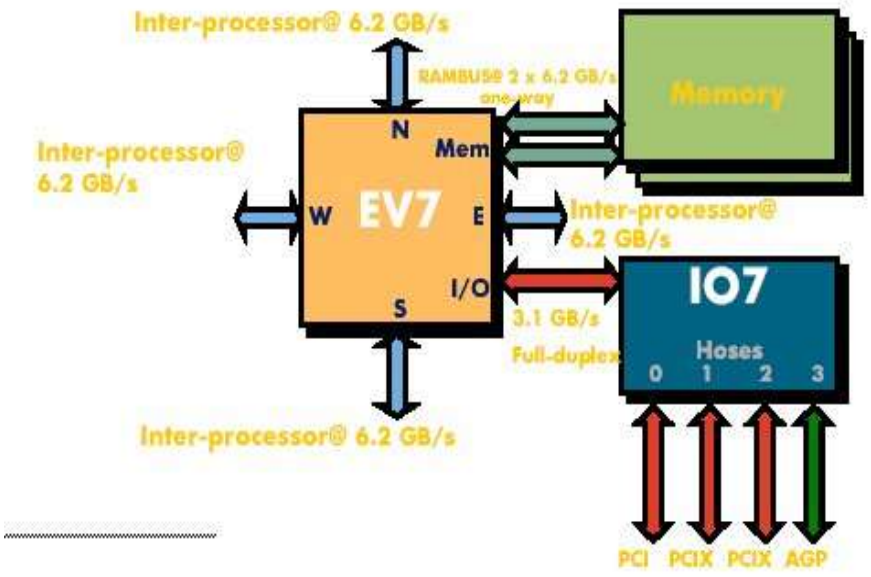
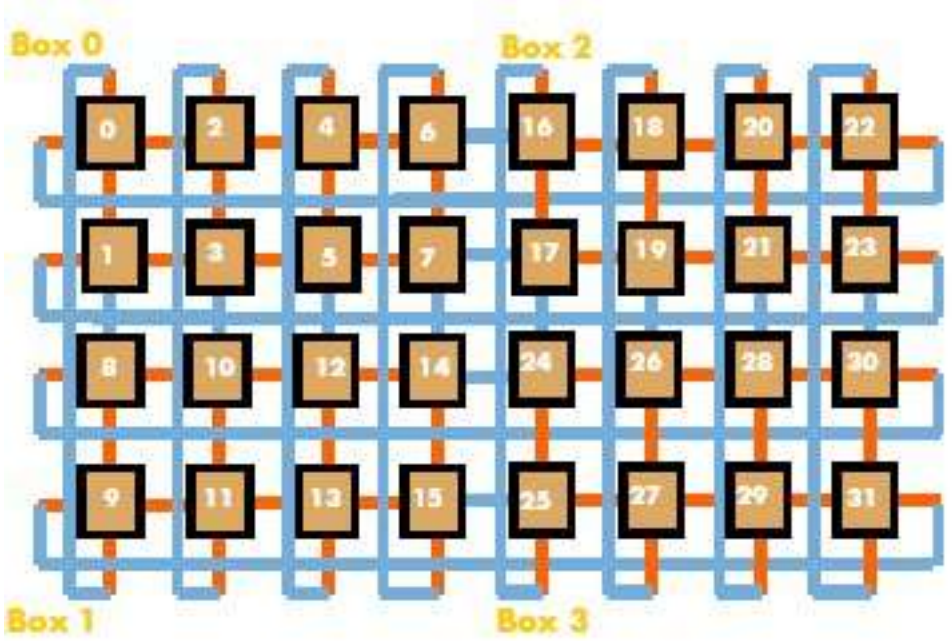


Figure 1a. EV7 interface to memory and I/O.

Z.Cvetanovic, GS1280, white paper, 2003

- EV7-alpha-CPU 1.15GHz
- Verbindungen zu 4 Nachbar-CPUs je 6.2GB/s !!!
- 2 Speicherkanäle mit je 6.2GB/s zum lokalen RAM !

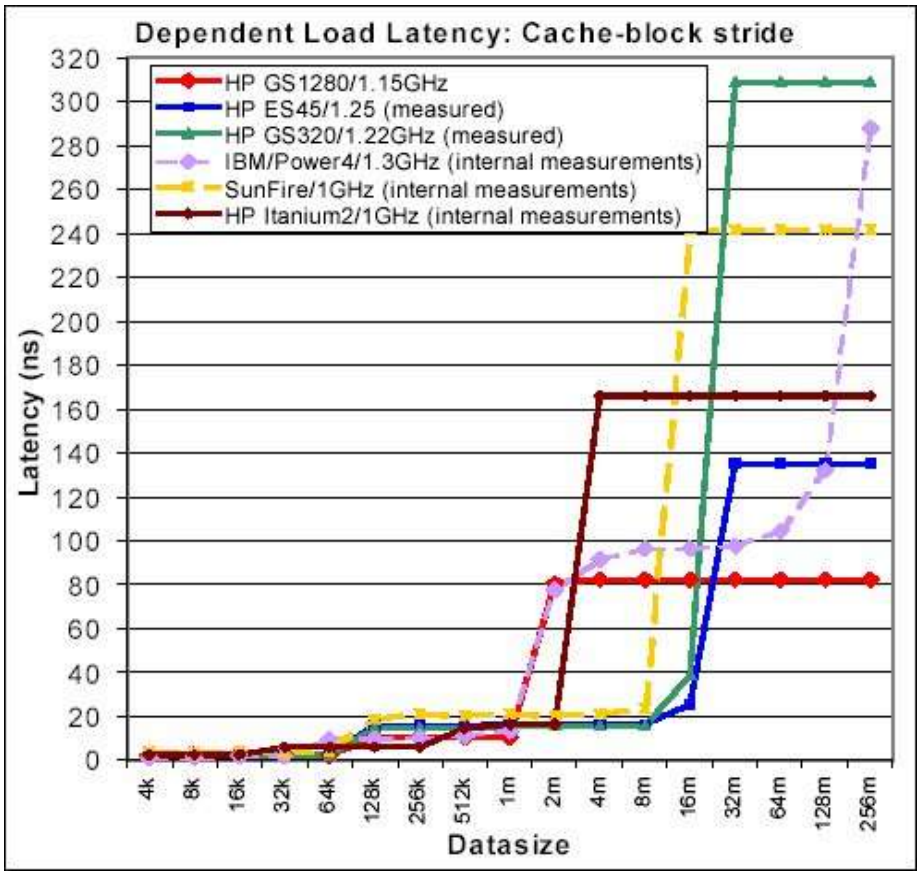
128GB Shared Memory, 32 CPUs



- Non Uniform Memory Architecture (NUMA)
- 4GB lokaler Speicher pro CPU
- aber auch schneller Zugriff auf entfernten Speicher (!!!)

Z.Cvetanovic, GS1280, white paper, 2003

lokaler Speicher



Z.Cvetanovic, GS1280, white paper, 2003

- Spitzenwerte
- aber PC konkurenzfähig (Preis/Leistung)
- PC hat inzwischen schnellere CPUs

Entfernter Speicher (remote)

82	141	177	149
133	169	216	176
176	215	250	216
150	184	228	192

Figure 13. Memory latencies (ns) on GS1280 (each square represents a CPU in a 16-CPU torus).

- Latenz unschlagbar!
- PC-Kluster völlig außer Konkurenz
- Wichtig bei wahllosen Zugriff über große Speicherbereiche.

Z.Cvetanovic, GS1280, white paper, 2003

eigene Messungen

write 1GB memset(buf,0,1024*1024*1024)

	1st loop	2nd loop	CPU-freq.	Speed
ES45	2383..2466ms	649..666ms	1250MHz	1.5GB/s
GS1280.locRAD	1966..1983ms	616..633ms	1150MHz	1.6GB/s
GS1280.remRAD	2266..3016ms	666..783ms		1.5GB/s
GS320.locRAD	5483..5516ms	2783..2949ms	731MHz	360MB/s
GS320.remRAD	8966..9866ms	4083..4166ms		245MB/s
PC-3.2GHz	2600ms	880..890ms (8*128MB)		1.2GB/s*

2^3 steps (sizeof long)

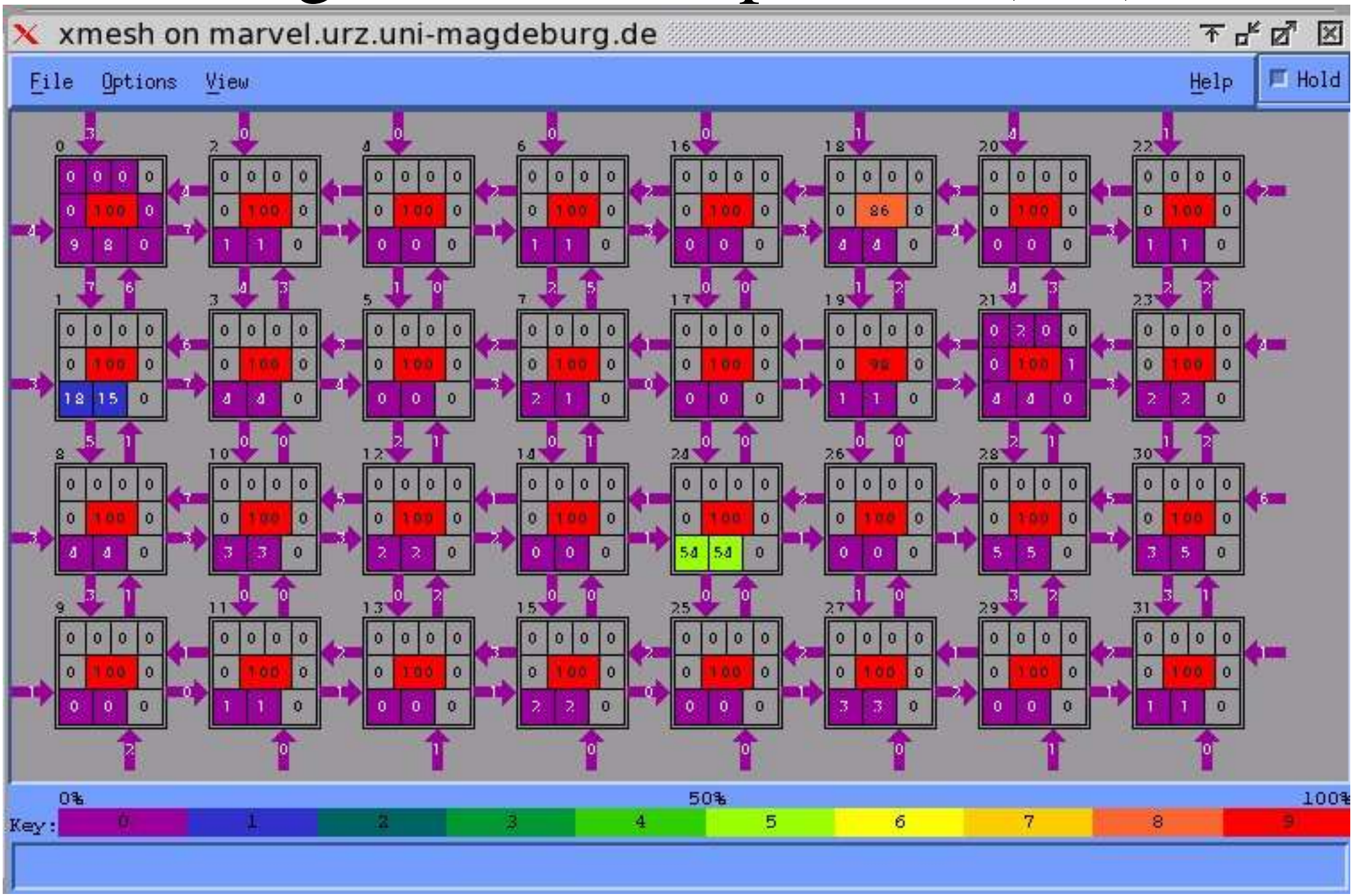
1e2*2^30B read-loop 21.8s 4697MB/s 1.7ns/long

2^13 steps (sizeof page, 8k), 6GB initialisiert (4GB local, 2GB remote)

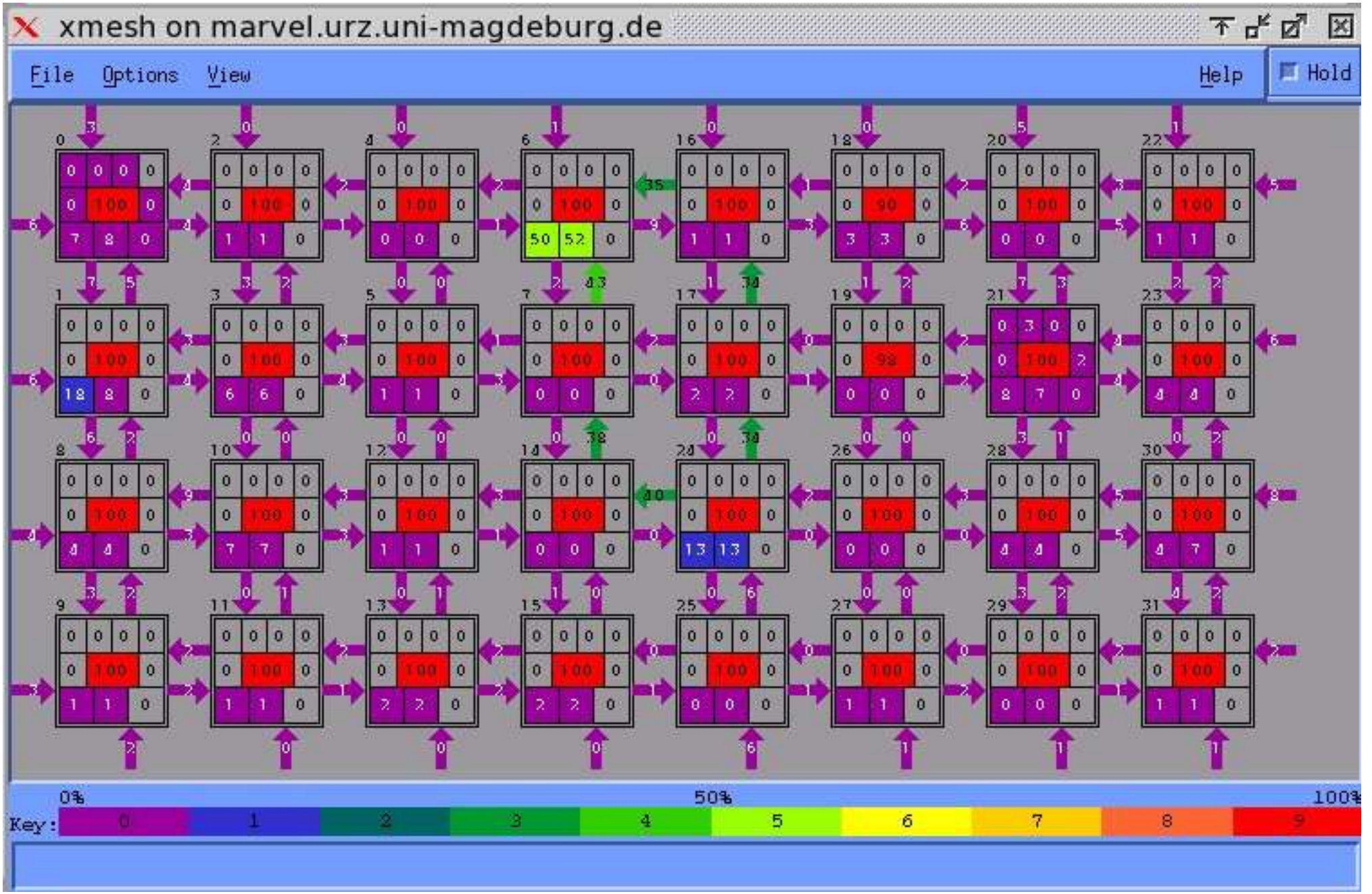
1st GB 145ns/long, 6th GB 239ns/long

* als “repz stos” kodiert (loop 570MB/s)

Zugriff lokaler Speicher (r24)



Zugriff remote Speicher (r24-r6)



Warum ist die Marvel so schlecht?
-- Software --

- Anpassung von Tru64 (OSF1) an NUMA
- Trick: Resource Affinity Domains (RADs)
- logische Teilung des Systems
- 1 RAD = 1 CPU + 4GB lokaler Speicher
- Ok, dient ja nur der Performance ...
- Achtung: RADs haben eigene VM-Manager,
  ---> Nebenwirkungen?

Virtueller Memory (VM)

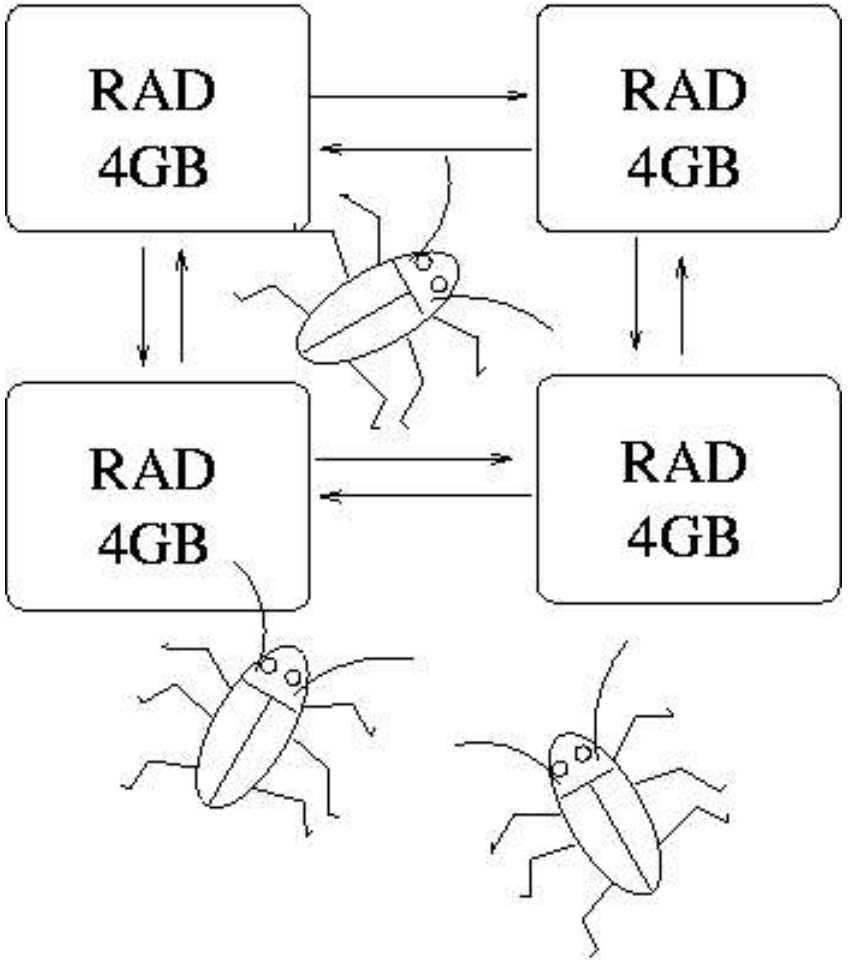
- 8k “Speicherhäppchen” (pages)
- Listen: aktiv, inactiv, free
- unified buffer cache (UBC)
- Tuning-Parameter (man sys_attrs_vm)
- Gibts auch in anderen Unixen.

VM und RADs

- Jedes RAD hat seinen eigenen VM-Manager (schlecht dokumentiert, versteckte Parameter).
- Unsichtbar für NonNUMA-aware Software
- Warum? RADs können Speicher von anderen RADs stehlen.
- Theoretisch ist also alles gut.

Eigentlich ok, aber ...

- positiv: lokaler Speicher bevorzugt (schnell)
- negativ: “shared memory” Verhalten nicht richtig umgesetzt (fehlerhaft)



Wie sollte es sein?

- Lokalen Speicher füllen (ca. 6GB/s, 80ns)
- Lokalen UBC reduzieren.
- Remote Speicher füllen (ca. 6GB/s, 140-250ns)
- Remote UBC reduzieren.
- Wenn kein Speicher mehr frei, pagen.
  - ca. 100x langsamer! (71MB/s)
  - Deshalb: Bei freiem Speicher niemals pagen!

Was passiert?

- Je nach Anwendung pagen nach wenigen Tagen!
- Swap wächst trotz vielen GB freien Speichers!
- Marvel wird extrem langsam!
- Warum?
  - Pages stehlen ging ab 14GB schief (behoben)
  - remote UBC wird nicht freigegeben (HP gemeldet)
- Problem auch bei anderen Kunden (Forum).

Erfahrungen mit dem Softwaresupport.

- Eine Katastrophe!
- Warum?
  - Strategie abblocken. Kein Bug. (6/2003)
  - Ändere doch Dein Programm, lieber Kunde (3/2005).
  - NUMA-Esoterik, karge Antworten, Level 3
  - Forum, Pathologisch? Huch, doch ein Bug.
  - Störe meine Kreise nicht! -- Experten unter sich.
  - Forum: lazy versus eager swap mode

Notlösung.

- Eigene Lösung: minimaler UBC (max. 1%)
- Nachteil: kein File cashing (slow I/O)
- seitdem stabil, kein unerklärliches pagen
  - Störungen durch Supportteam und I/O-Applikationen
  - Erstmalig nutzen (normale) Applikationen bis zu 100GB ohne Performance-Verlust.

Gibts Hoffnung?

- Mehr Druck!
  - Internet und Foren (done, wirkungslos?)
  - Vortrag im Arbeitskreis HPC.
  - Was meint die Konkurrenz?
  - Nie wieder HP-Software?
- HPs Support besuchen/schulen?
- Juristische Schritte? Gewährleistung.