New hardware at Mainz

High performance 12-core PC Dell R710

Two X5670 CPUs (24 virtual CPUs) @ 2.93GHz (3.33GHz turbo)
24GB memory @ 1333MHz
6*500GB with a fast PERC H700 ($\approx 450\frac{MB}{s}$ serial writing)
Special ATLAS conditions: 3500€

Low power 12-core PC Dell R710

Two L5640 CPUs (24 virtual CPUs) @ 2.26GHz (2.8GHz turbo)
24GB memory @ 1333MHz
6*500GB with a fast PERC H700 ($\approx 450\frac{MB}{s}$ serial writing)
Special ATLAS conditions: 3200€

48 x 1G-port switch Dell PowerConnect 6248

two SFP+ modules (two ports each)
Special ATLAS conditions: 888€ + 100€

New hardware at Mainz

5 singlecores
2*1G links
1*100M for netboot

2 * 12 cores
2*10G links
4*1G links
2/1*1G links for management
2TB diskspace each

2 * 4cores
4*1G links
1*1G link for netboot

1*4 cores
1*10G link
1*1G link for netboot

Power consumption with different power supplies

Power supply:	1x570	2x570	1x870	2x870
low load	156 ± 3	163 ± 3	171 ± 3	**190**± 4
high load	324 ± 6	332 ± 7	329 ± 6	**351**± 7

Table: Power consumption of the high performance PC in Watts

Power supply:	1x570	2x570	1x870	2x870
low load	101 ± 2	**113**± 2	114 ± 2	138 ± 2
high load	249 ± 4	**252**± 4	250 ± 4	273 ± 4

Table: Power consumption of the low power PC in Watts

High load via linpack (10k equations) benchmarks (HT on) taking 19GB memory:
Fast PC: 102 GFLOPS (**292** $\frac{MFLOPS}{Watt}$)
Low Power PC: 82 GFLOPS (**337** $\frac{MFLOPS}{Watt}$)

Let's talk about racks

10kW racks are suggested so far

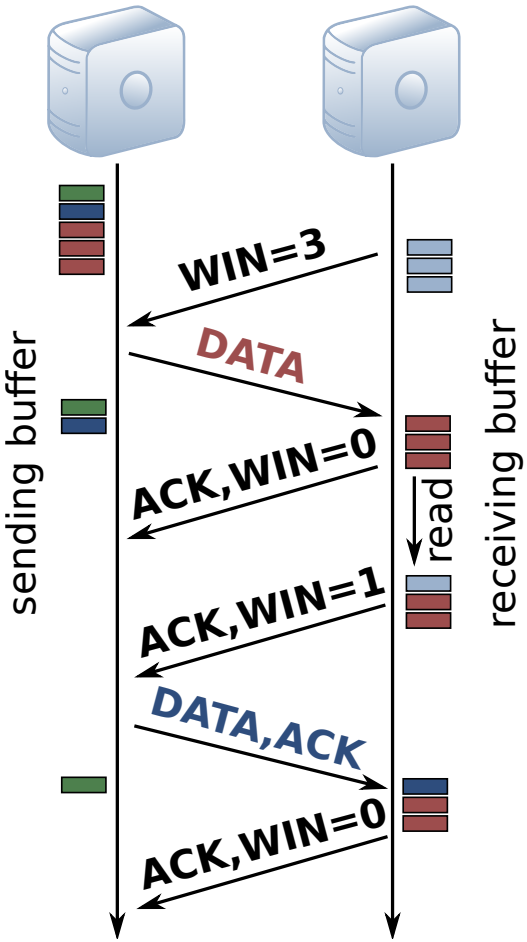
351 Watts per PC ⇒ up to 28 PCs per rack

2U-PCs fit without a problem!

If you really want to buy expensive 2U-PCs I would put disks into every PC instead of building a separate disk farm (Dell R510: up to 12 x 3,5" disks per PC) → need 100PCs to achieve 300TB. Better: Buy 1U-PCs and external raid arrays! (also if you want a separate farm)

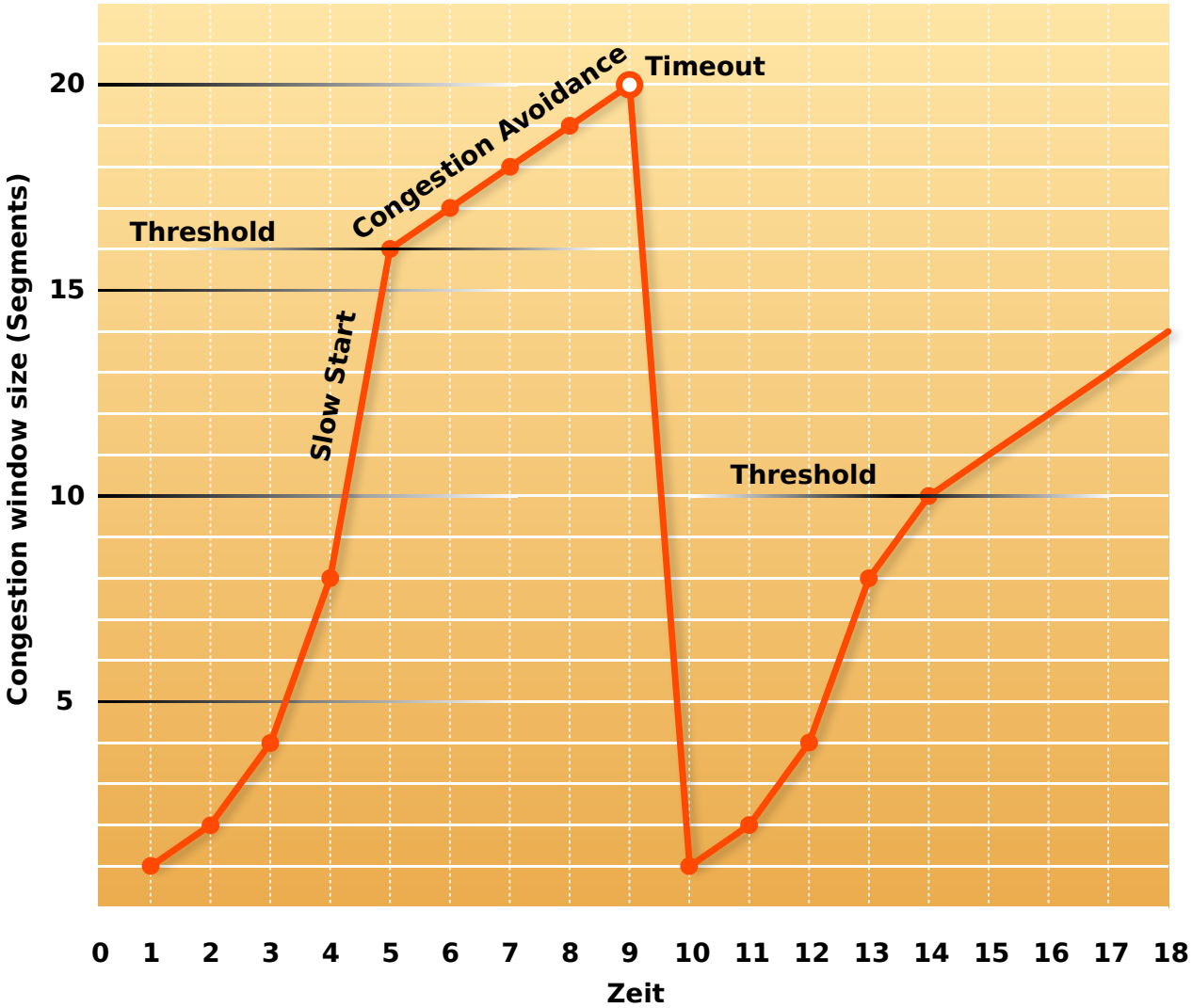
Flow control

Sliding window

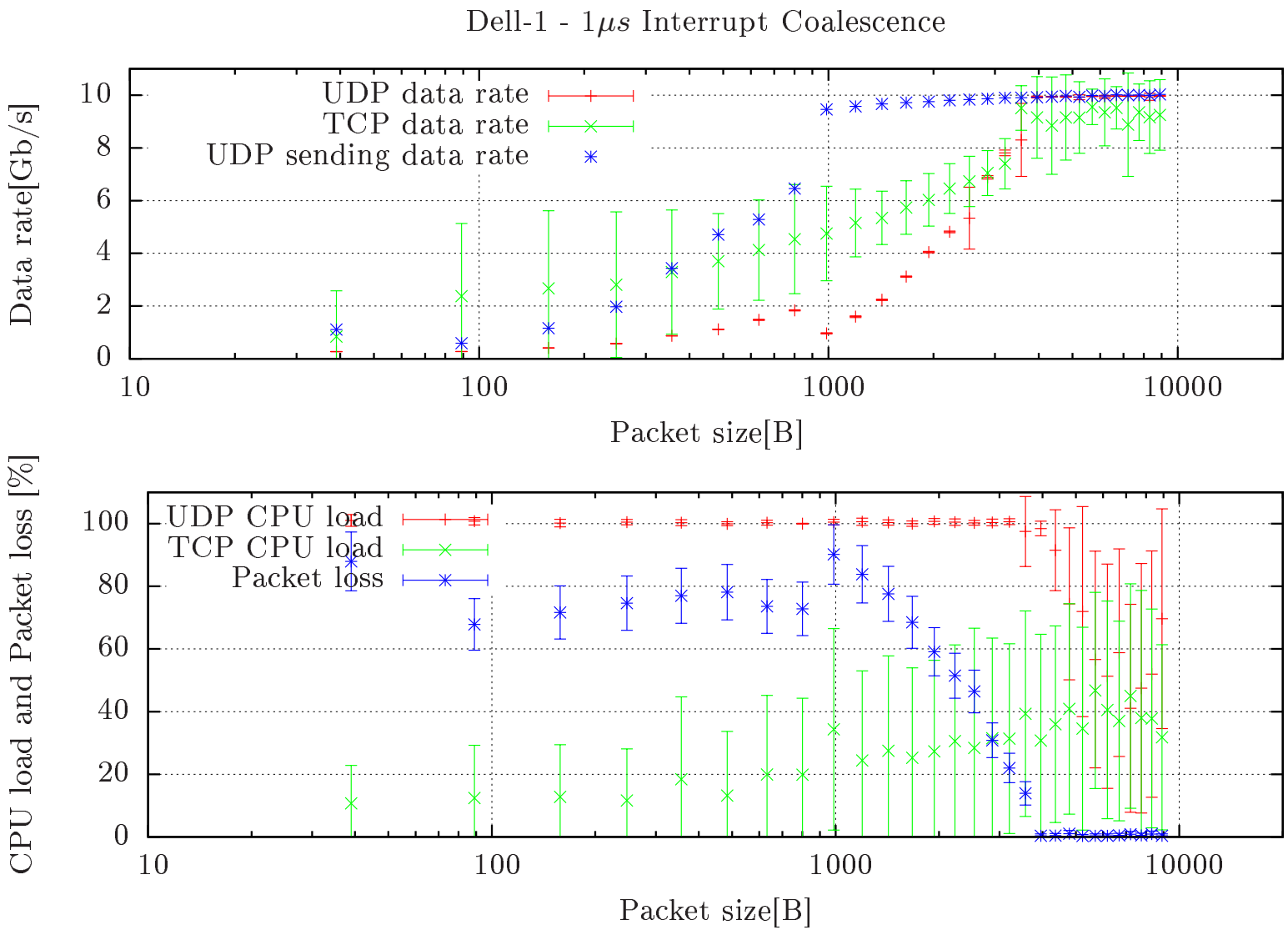


Congestion Avoidance

Congestion window

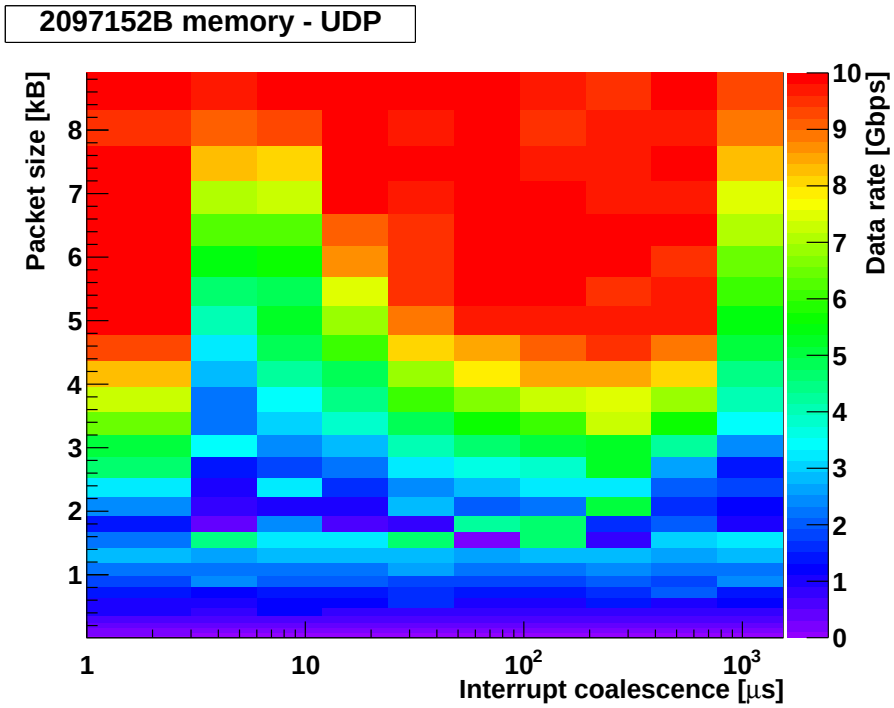


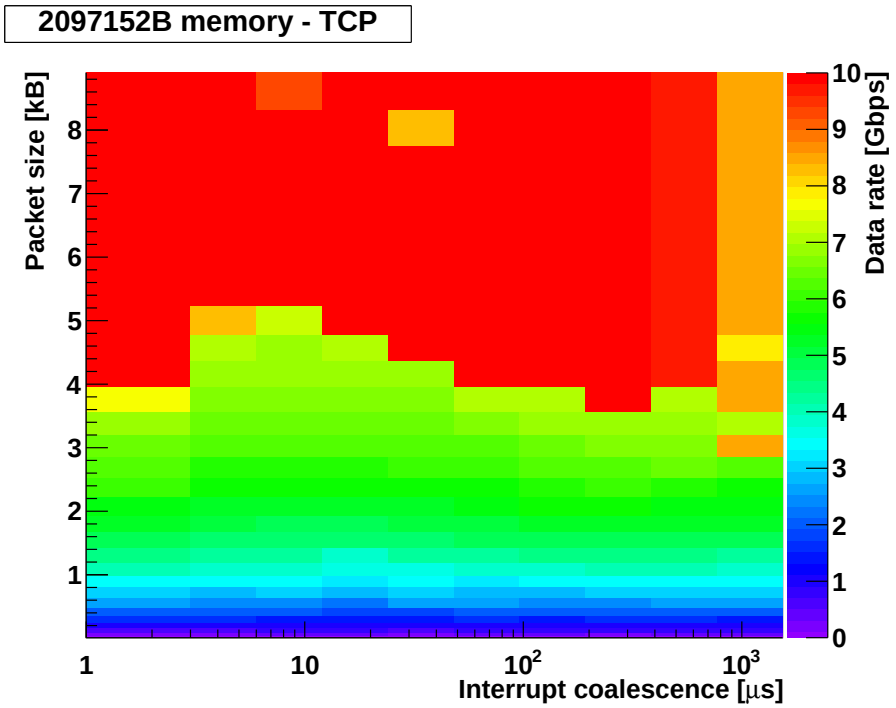
Packet sizes



Interrupt coalescence
Data rate

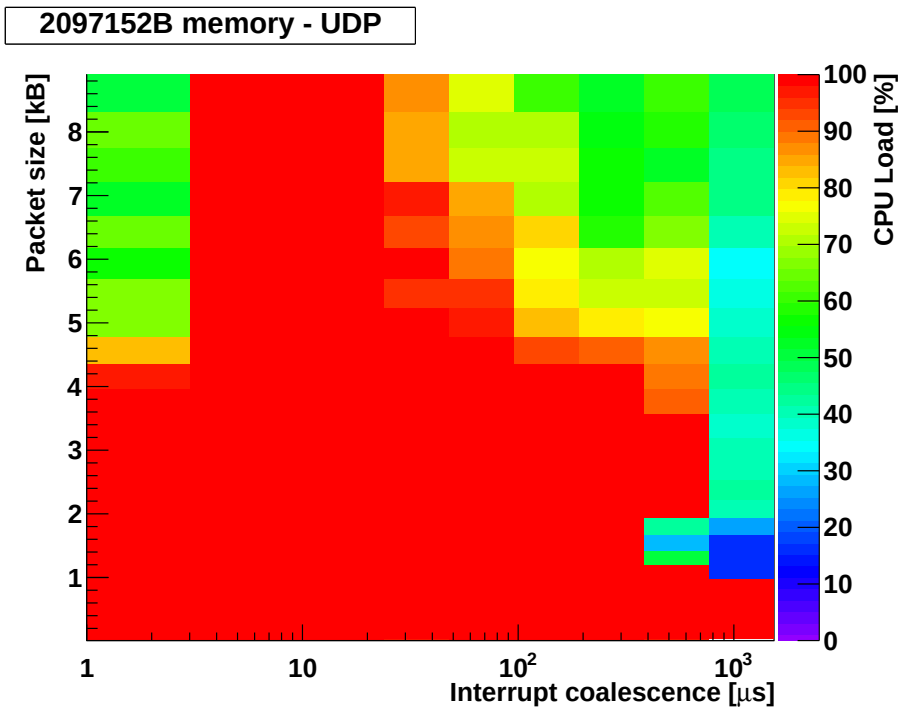
$1\mu s \hat{=}$ automatic calibration of interrupt coalescence

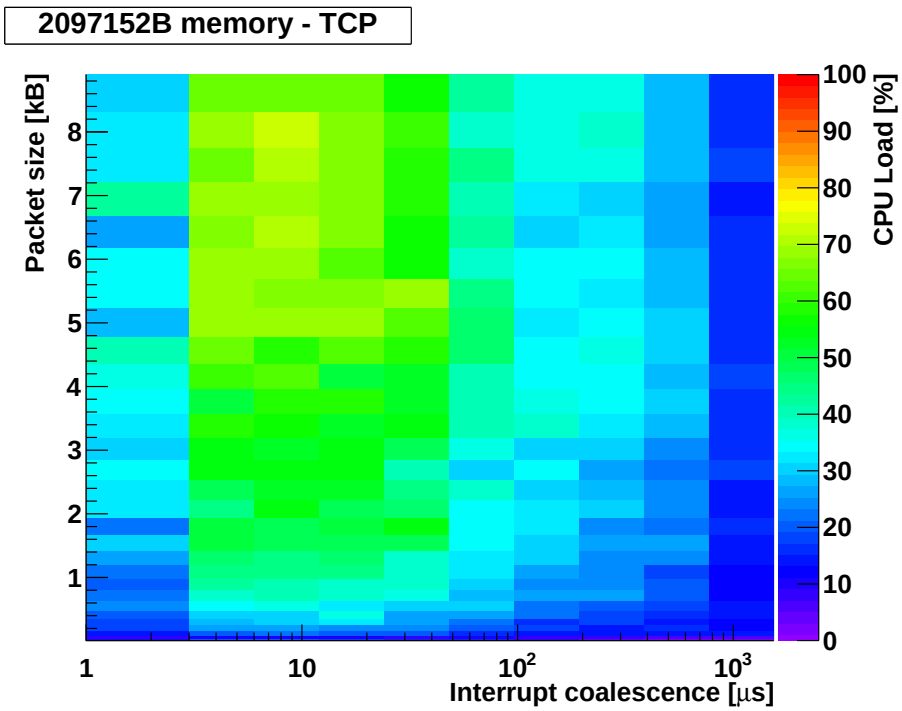




Interrupt Coalescence
CPU load

$1\mu s \hat{=}$ automatic calibration of interrupt coalescence





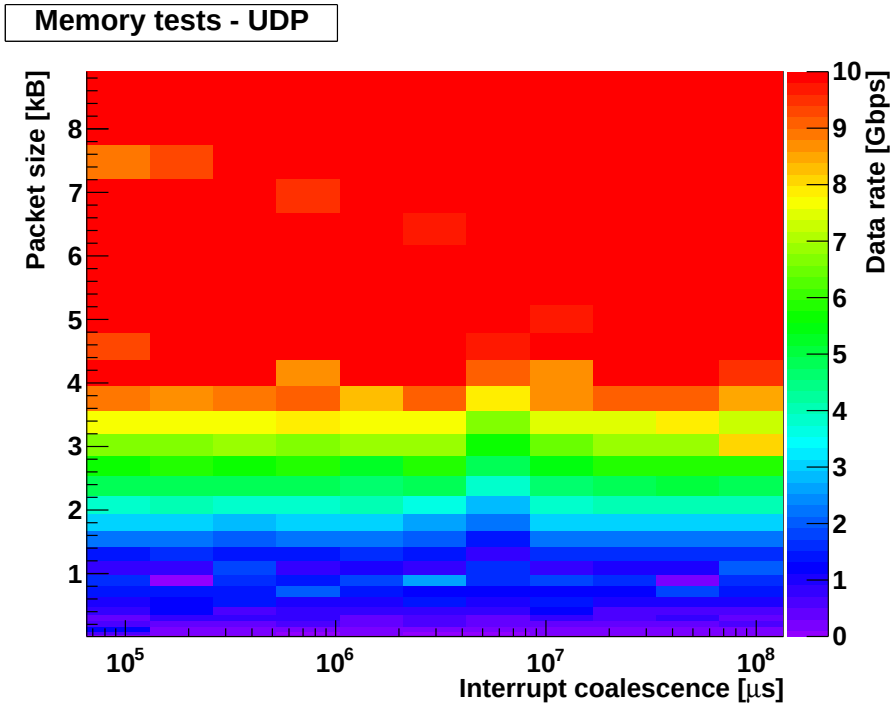
Window sizes

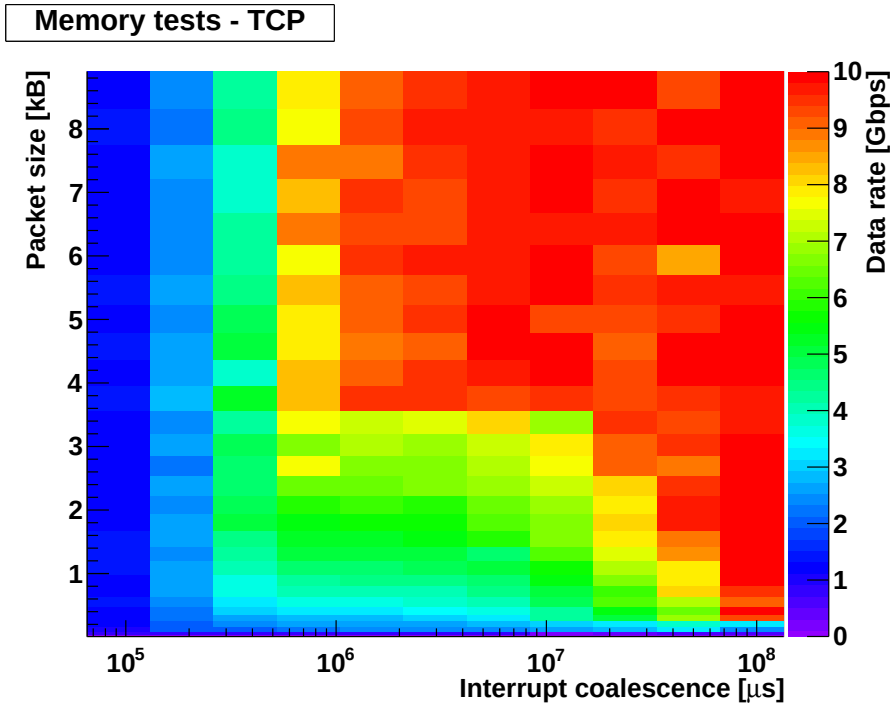
I've run several tests with different memory settings and a interrupt coalescence of $1\mu s$:

Settings

```
sysctl -w "net.ipv4.tcp_rmem=$i $i $i"
sysctl -w "net.ipv4.tcp_wmem=$i $i $i"
sysctl -w "net.core.rmem_max=$i"
sysctl -w "net.core.wmem_max=$i"
sysctl -w "net.ipv4.udp_mem=$i $i $i"
sysctl -w "net.ipv4.udp_rmem_min=$i"
sysctl -w "net.ipv4.udp_wmem_min=$i"
```

Window sizes





The second bin is the default value

Conclusions

We should use TCP because:

- Flow control and congestion avoidance
- TCP-optimized hardware (offload etc.)
- Adding reliability on top of UDP only in user mode (TCP: inside Kernel)!
- UDP: can only write MTU to socket; TCP: data stream instead
- Naggle algorithm helps you sending large packages (low CPU as shown)