Energy-Efficient Data Transfers in Radio Astronomy with Software UDP RDMA

Third Workshop on Innovating the Network for Data-Intensive Science, INDIS16

Przemek Lenkiewicz, Researcher@IBM Netherlands
Bernard Metzler, Researcher@IBM Zurich Research Lab
Chris Broekema, Researcher@ASTRON Netherlands Institute for RadioAstronomy

ASTRON Netherlands Institute for Radio Astronomy

Table of contents/Agenda template

The DOME project

IBM Netherlands

IBM Zürich Research Lab

ASTRON Netherlands Institute for Radio Astronomy

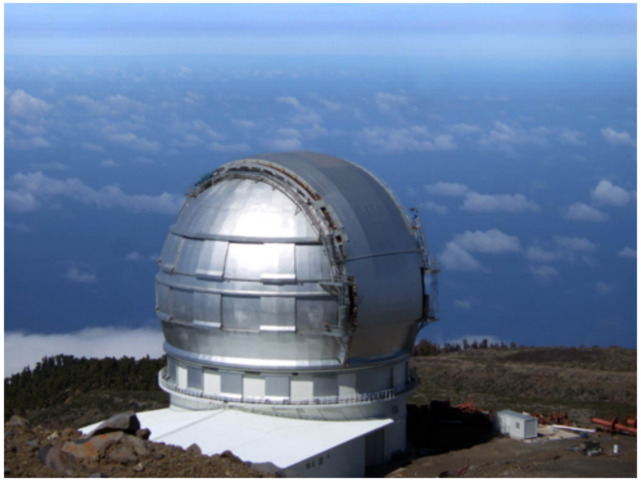
Radio astronomy & The Square Kilometre Array

A brief introduction

Astronomy

- Lenses, mirrors, sensors
- Light
- Picture of object

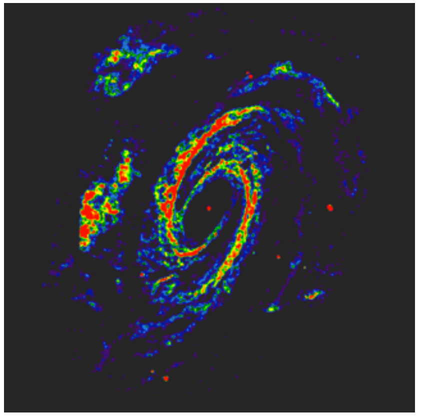
Gran Telescopio CANARIAS



The M33 Galaxy

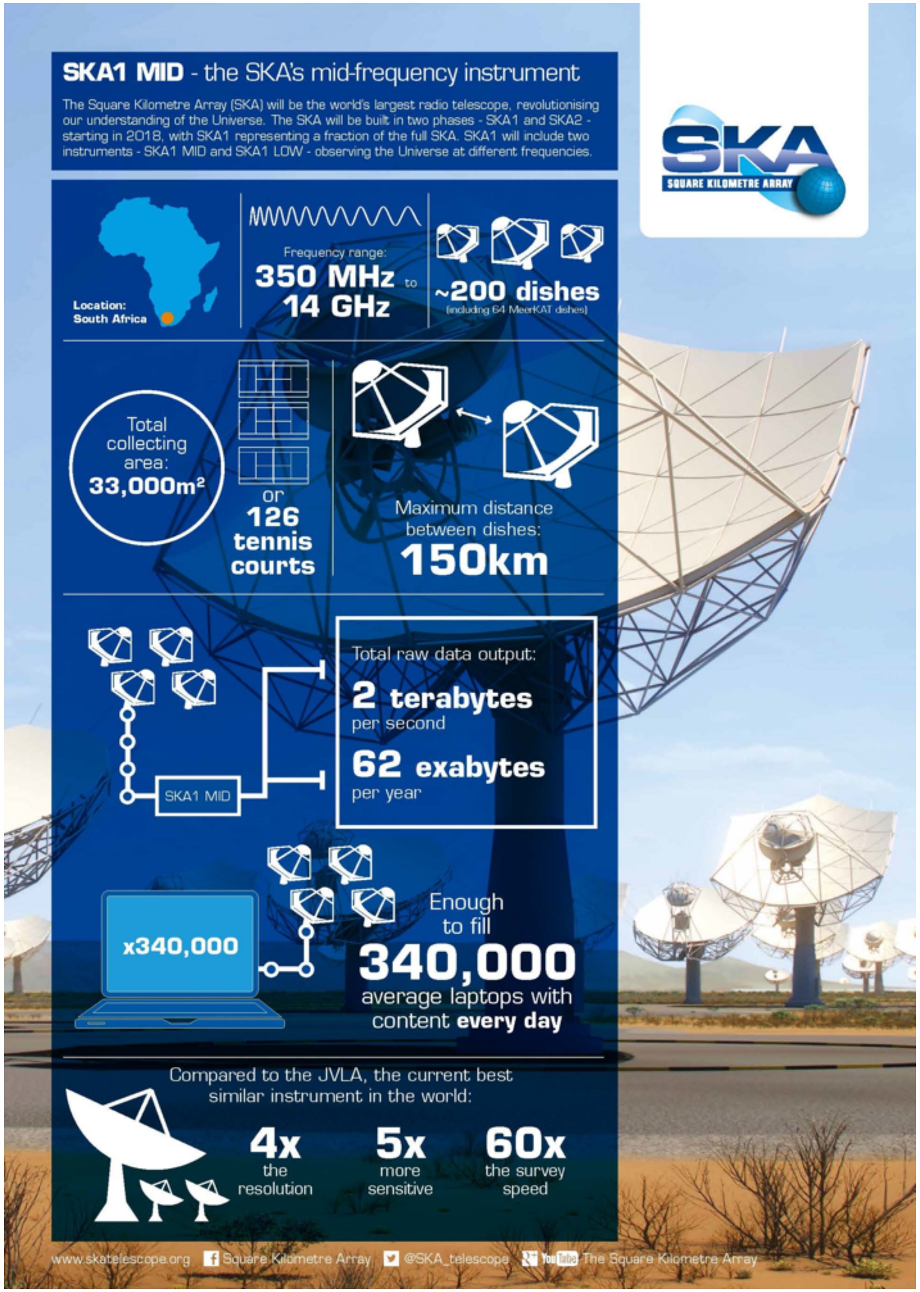
- Array of antennas and/or dishes
- Radio frequencies
- Map of radio sources

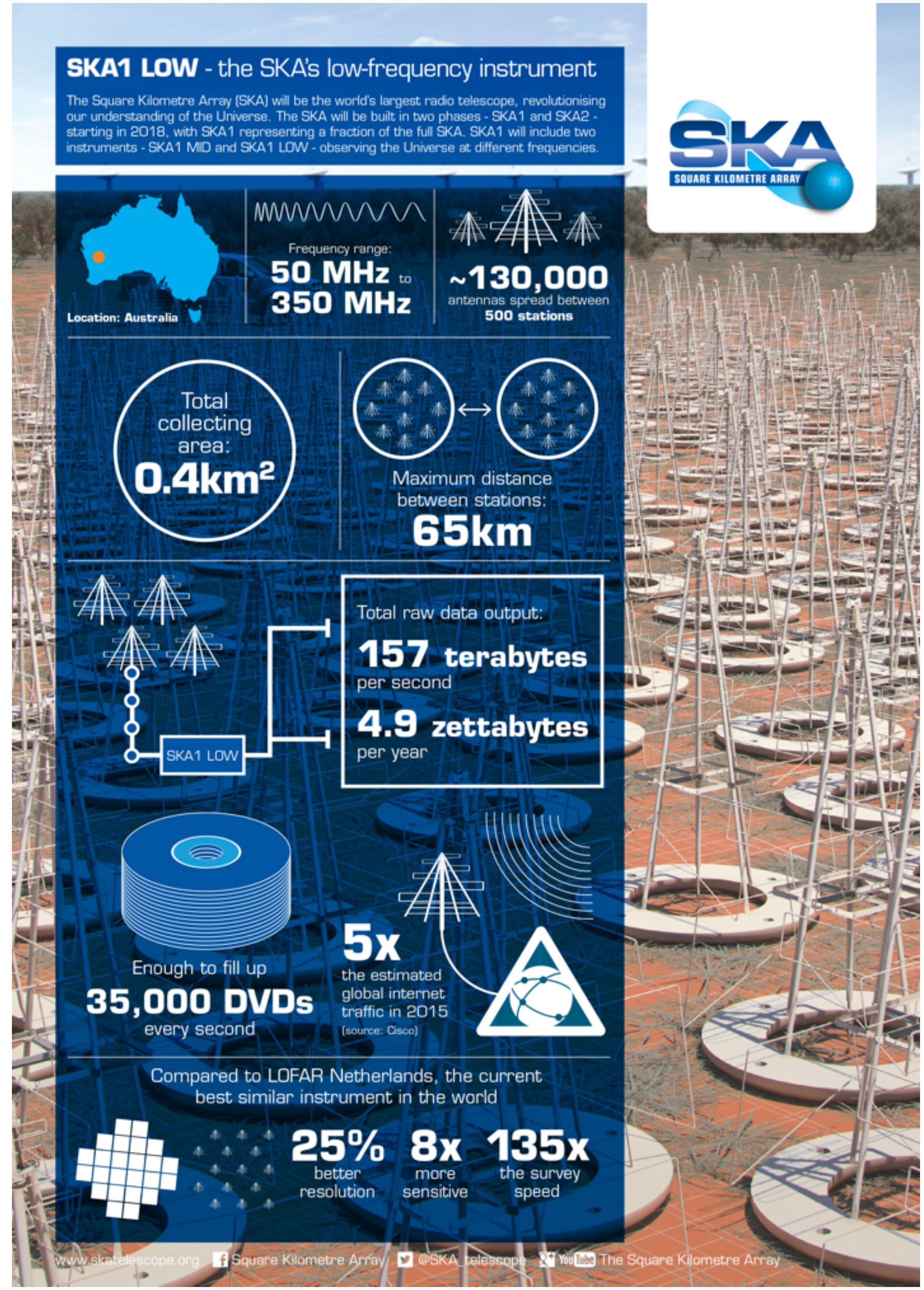
Low-Frequency Array (LOFAR)



The M81 Galaxy

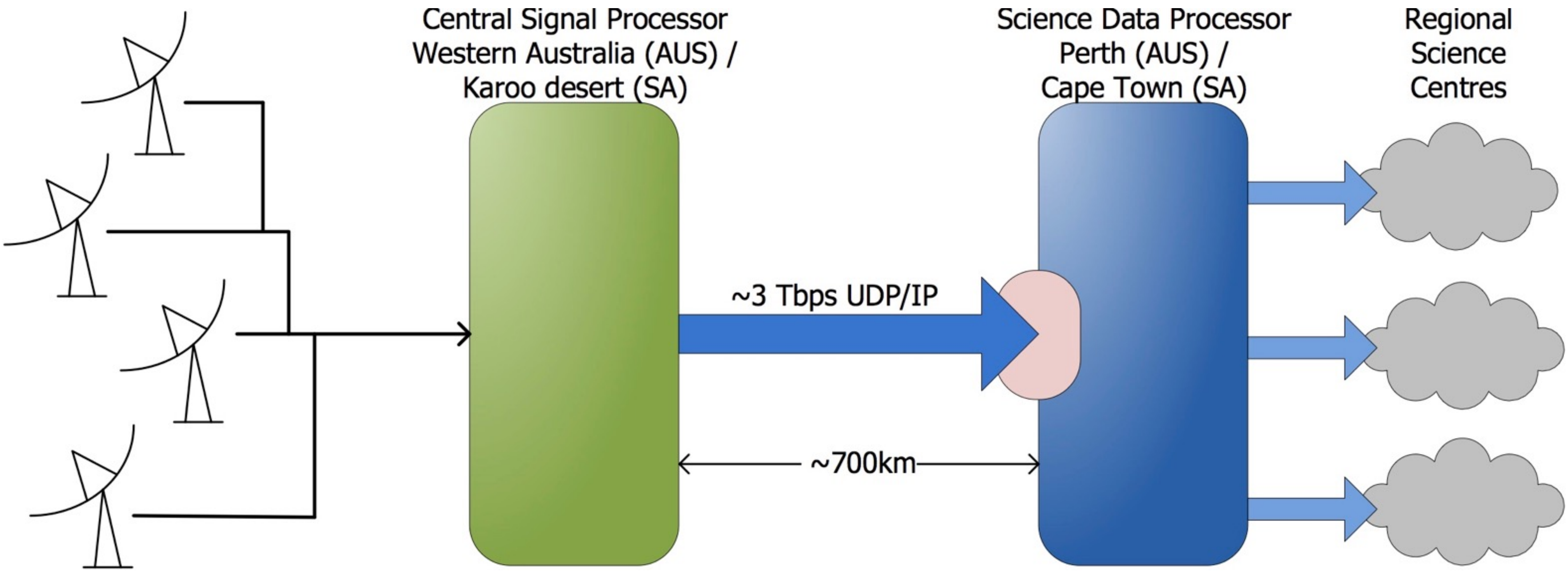
The Square Kilometre Array





Radio astronomy data transport

SKA telescope data flow



SKA Phase 1 in numbers
(italics are derived and/or speculative)

	SKA1 MID	**SKA1 LOW**
Location	Karoo, South Africa	Western Australia
Number of receivers	197 (133 SKA + 64 MeerKAT)	131.072 (512 st x 256 el)
Receiver diameter	15 m (13,5 m MeerKAT)	35 m (station)
Maximum baseline	150 km	65 km
Frequency channels	65.536	65.536
SDP input bandwidth	3,1 Tbps	3,1 Tbps
_Req’d Compute capacity_*	_20-72 PFLOPS_	_16-41,5 PFLOPS_
Archive growth rate	_10 – 100 Gbps (50yr life)_	_25 – 100 Gbps (50yr life)_
SDP Energy budget	_<5MW_	_<5MW_

*These are sustained PFLOPS, computational (in)efficiency not included

Cost cap for the first phase of SKA: € 650 M (2014)

Compare: Top 500 development

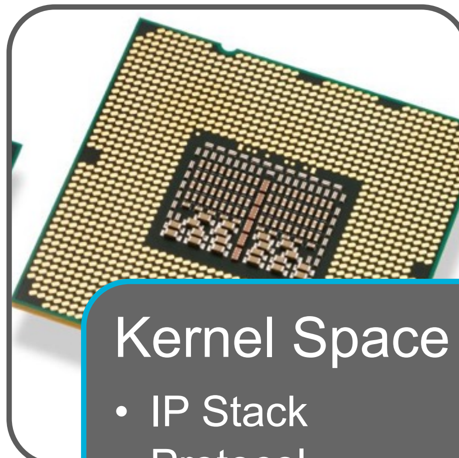
- Efficiency of radio astronomy algorithms: ~10%
- Effective required compute capacity: hundreds of PFLOPS

Souce: top 500 lists

Receive data through hierarchical OS

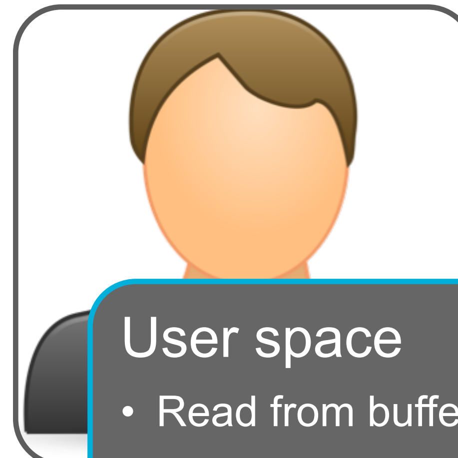
NIC

- Packet arrives
- DMA to kernel space
- Interrupt
- Context switch



Kernel Space

- IP Stack
- Protocol
- Memcpy to user space
- Context switch
- Interrupt



User space

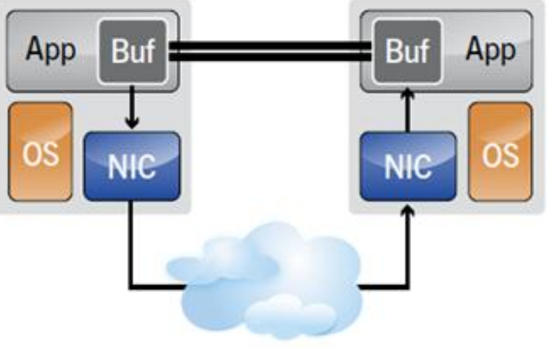
- Read from buffer

Requirements for astronomical data transport service

- Very high data rates – Terabits per second per instrument
- Almost entirely uni-directional traffic
- UDP/IP over Ethernet
- Prioritizing bandwidth over latency
- Desire for very high energy efficiency
  – Receiving end crucial!
- Full reliability is not crucial, some data loss is tolerable

Approach - RDMA

- Moving data from user space memory of one machine to that of another
- No involvement of host operating system
- Memory buffers registered with the local RDMA-capable network adapter (RNIC) and usually pinned to local physical memory
- Fully asynchronous to allow overlapping communication and computation



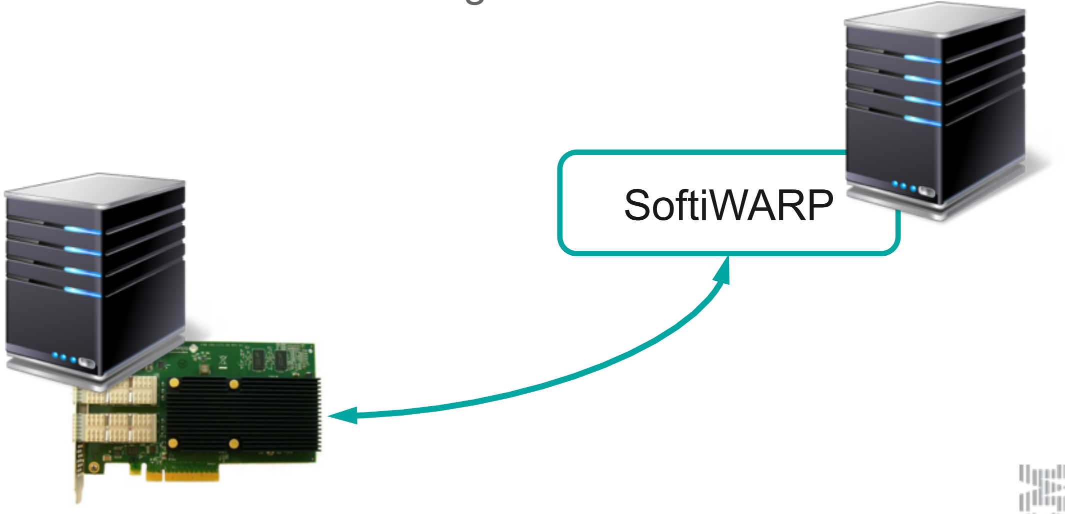
RDMA in radioastronomy

- We looked at:
  - RoCE
  - iWARP
  - Infiniband
- Reliable connection only
- Short range

Approach – SoftiWARP UDP

- SofiWARP (SIW)
- An open source software implementation of the iWARP protocol suite
- Developed at the IBM Zurich Research Lab and available from GitHub
  – https://github.com/zrlio/softiwarp
- Exports the OpenFabrics RDMA API to both user space and kernel space applications
- Fully compatible with hardware iWARP RNICs
- Utilizes kernel sockets for efficient communication and less data touching

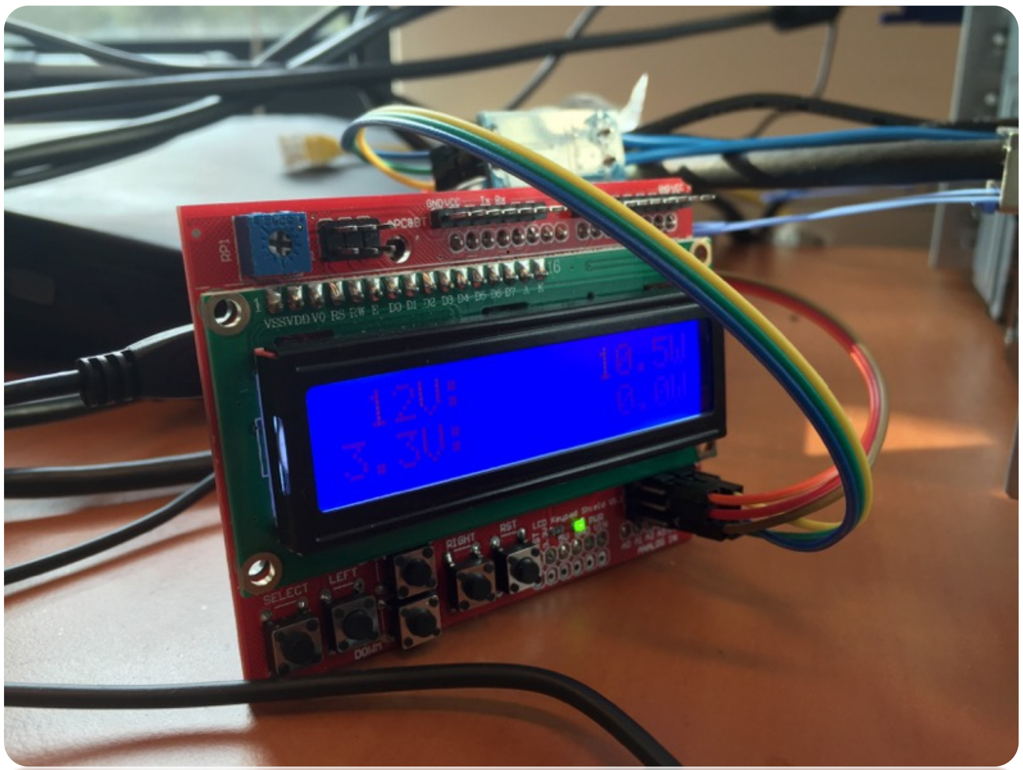
- In DOME:
  – Development of UDP transport layer

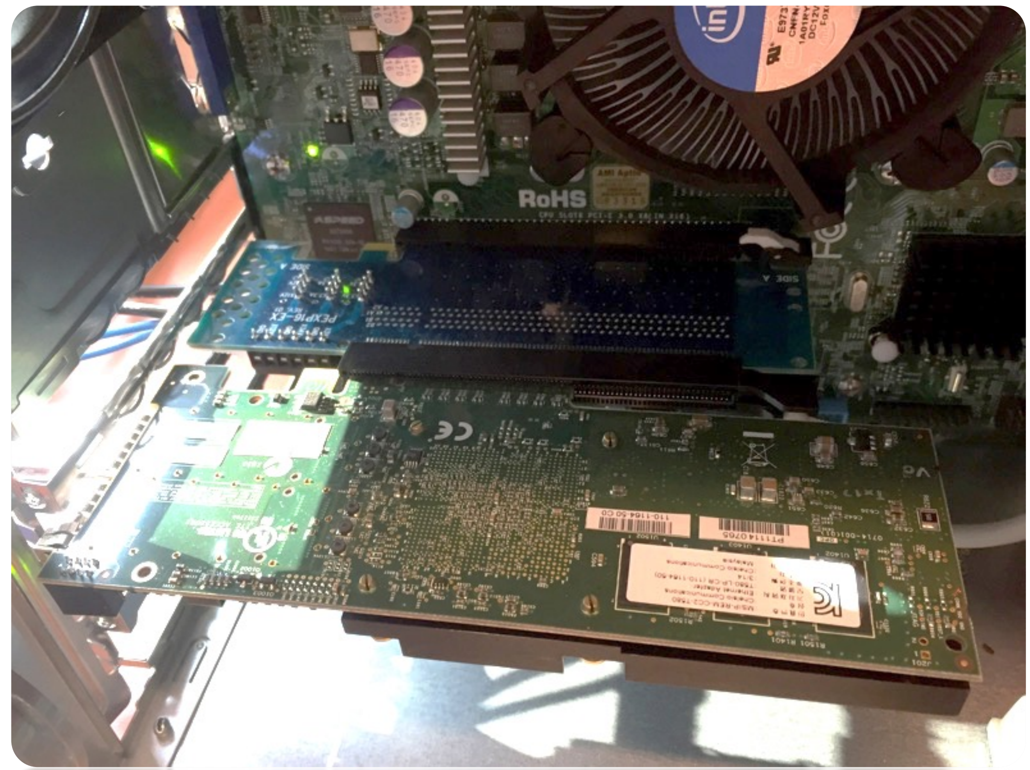


Experiments and measurements

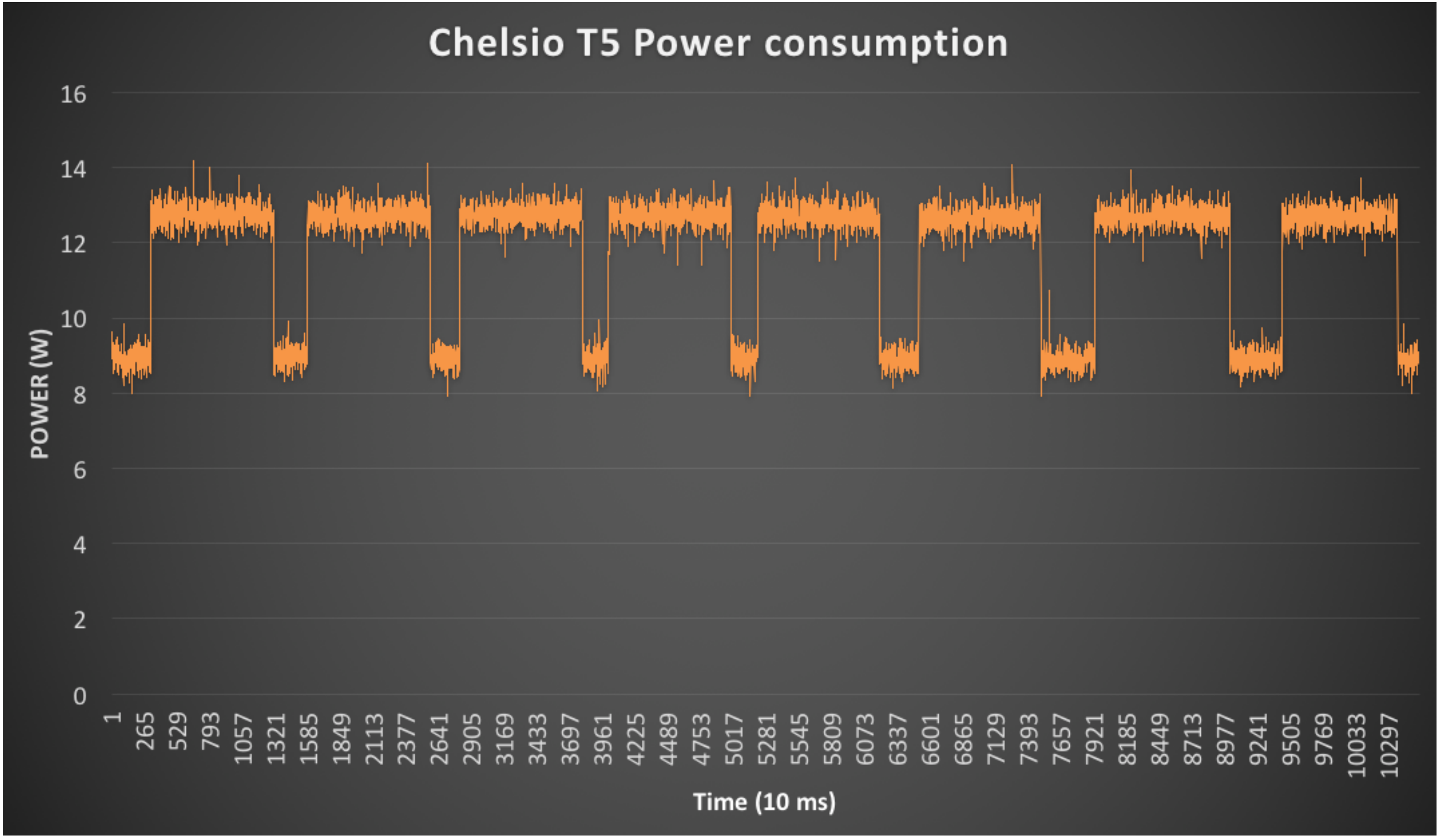
Power sensor for PCIe card slot

- ARDUINO board
- Current flow sensors
- PCIe riser card





Chelsio T5



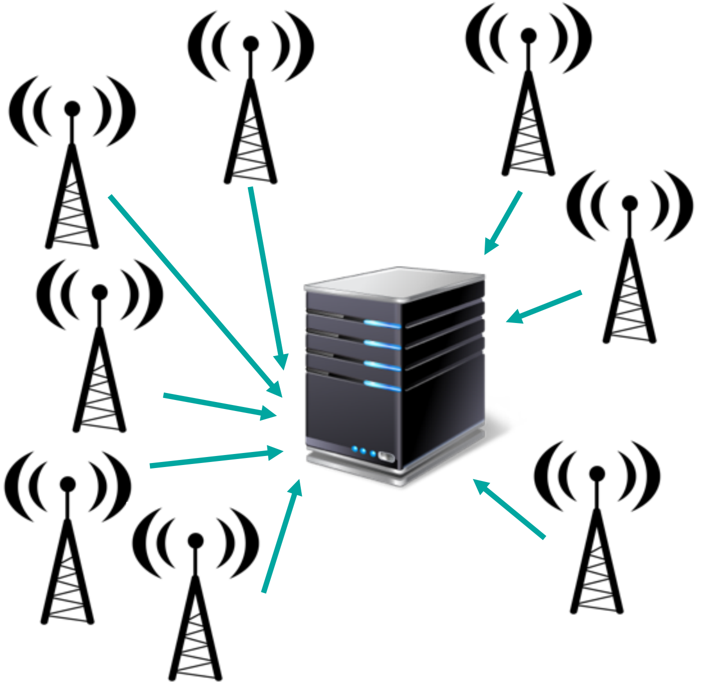
Chelsio T5 Power consumption
POWER (W)
16
14
12
10
8
6
4
2
0
1
265
529
793
1057
1321
1585
1849
2113
2377
2641
2905
3169
3433
3697
3961
4225
4489
4753
5017
5281
5545
5809
6073
6337
6601
6865
7129
7393
7657
7921
8185
8449
8713
8977
9241
9505
9769
10033
10297
Time (10 ms)

Power consumption measurements

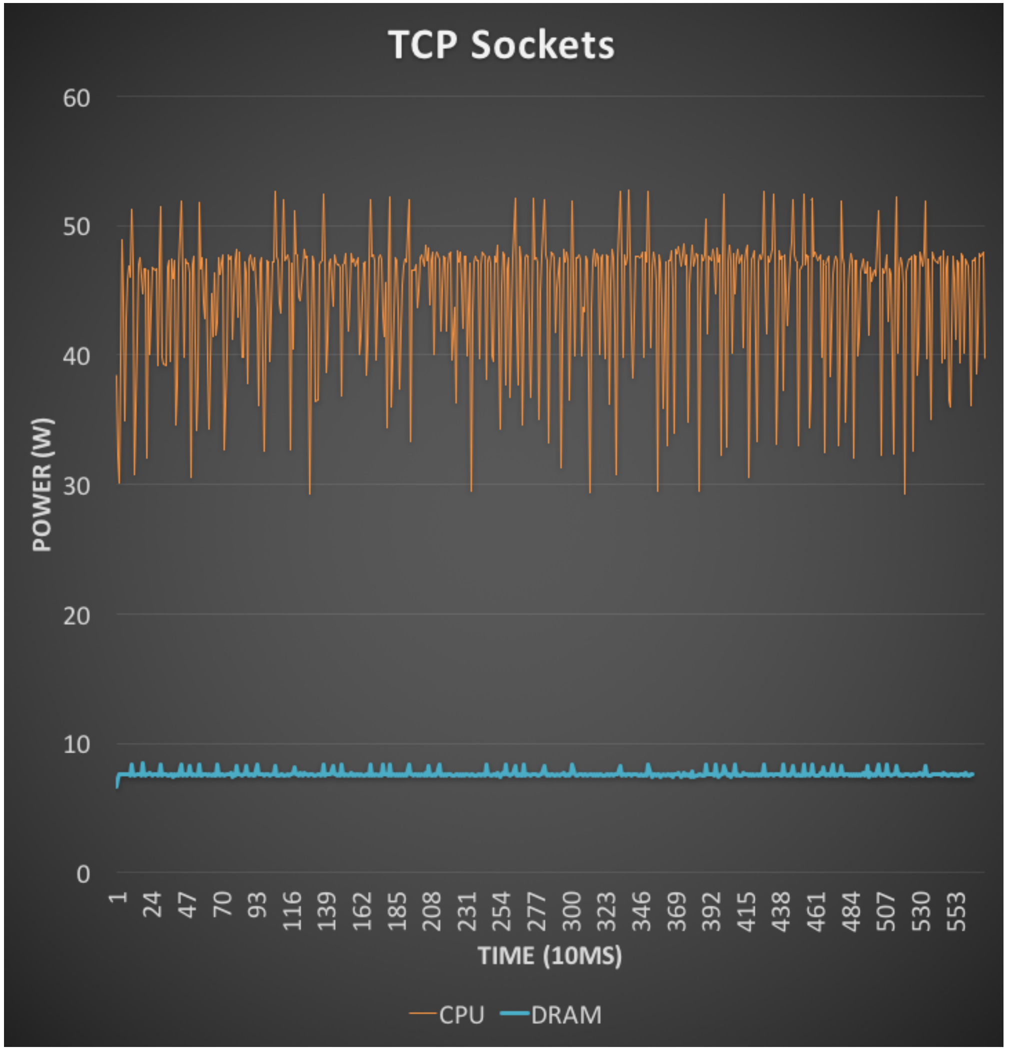
- CPU and DRAM power consumption
- RAPL for power readings
- Radio-astronomy traffic simulation
- Netperf benchmarking tool

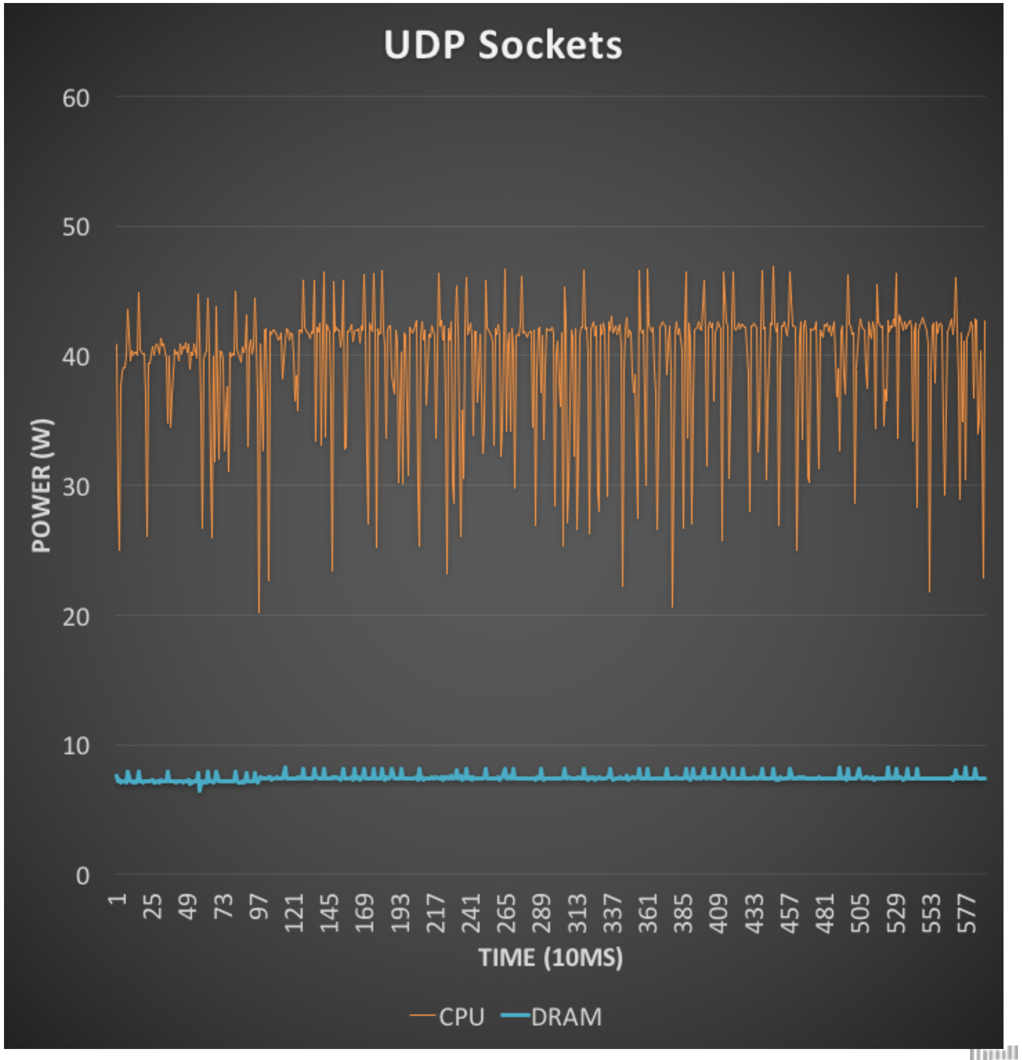
Radio-astronomy data flow

- Mimic the data flow from LOFAR
- Emulate the data produced by a LOFAR Remote Station Processing (RSP) board.
- UDP/IP data stream, ~760 Mb/s, packets of 8 kB.
- 50 Data streams received by a single CPU core
- Our emulator supports:
  – TCP Sockets
  – UDP Sockets
  – Softiwarp TCP
  – Softiwarp UDP

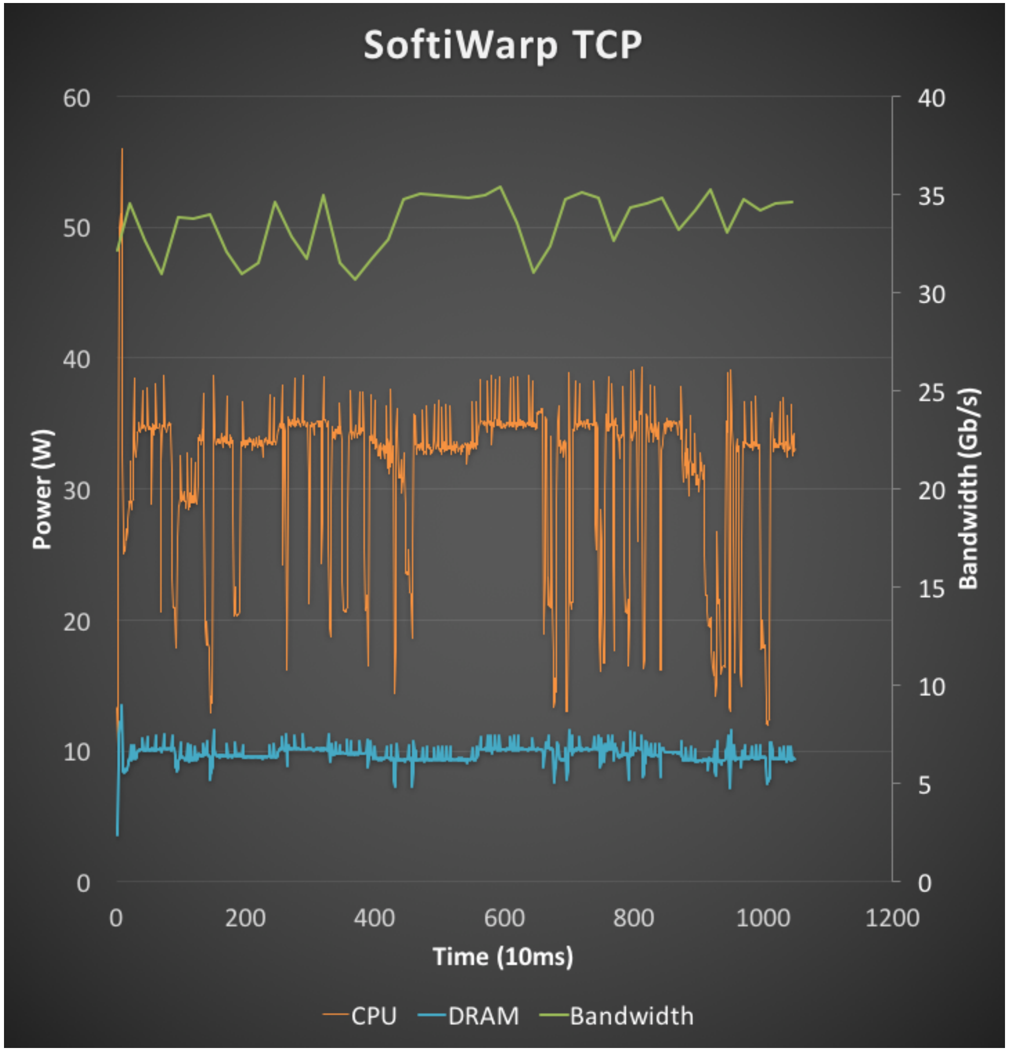


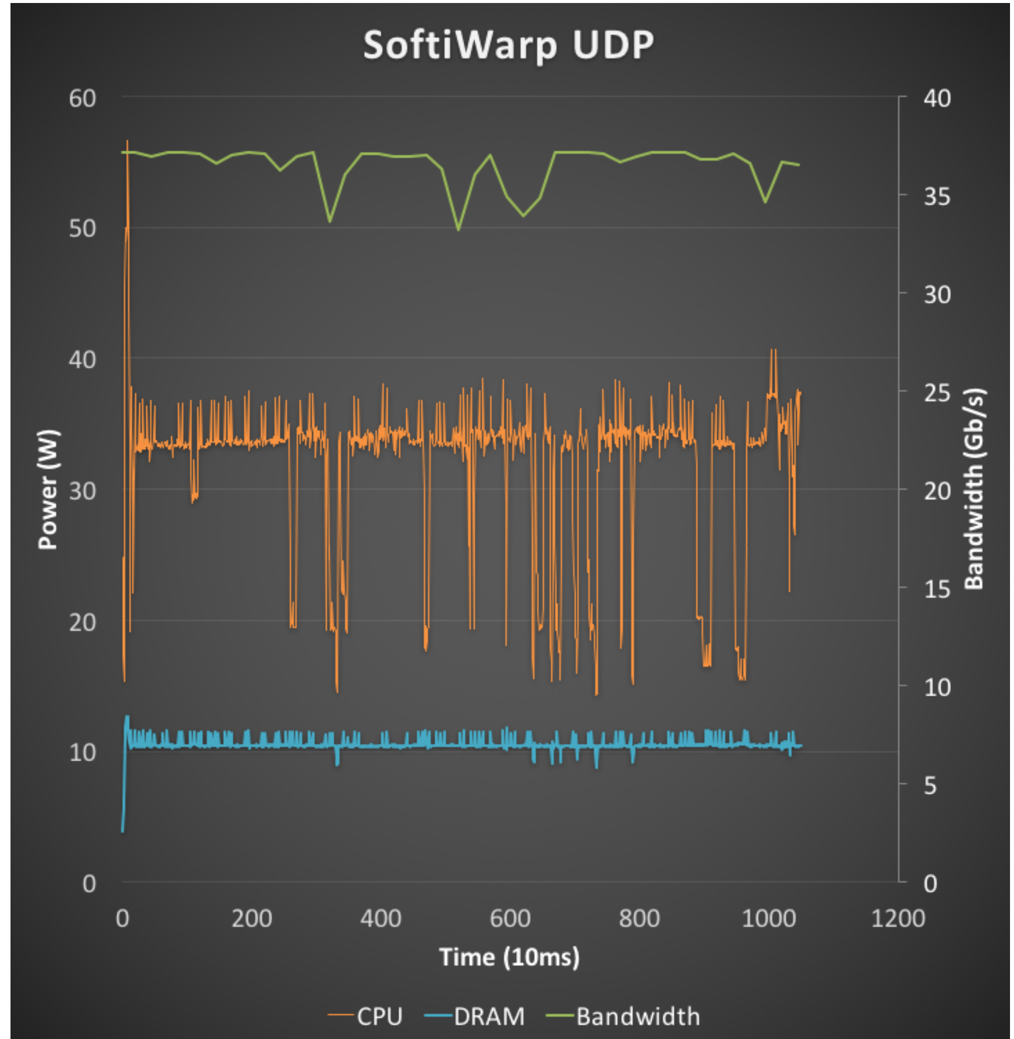
Radio-astronomy data flow





Radio-astronomy data flow





Power measurements, no offloading

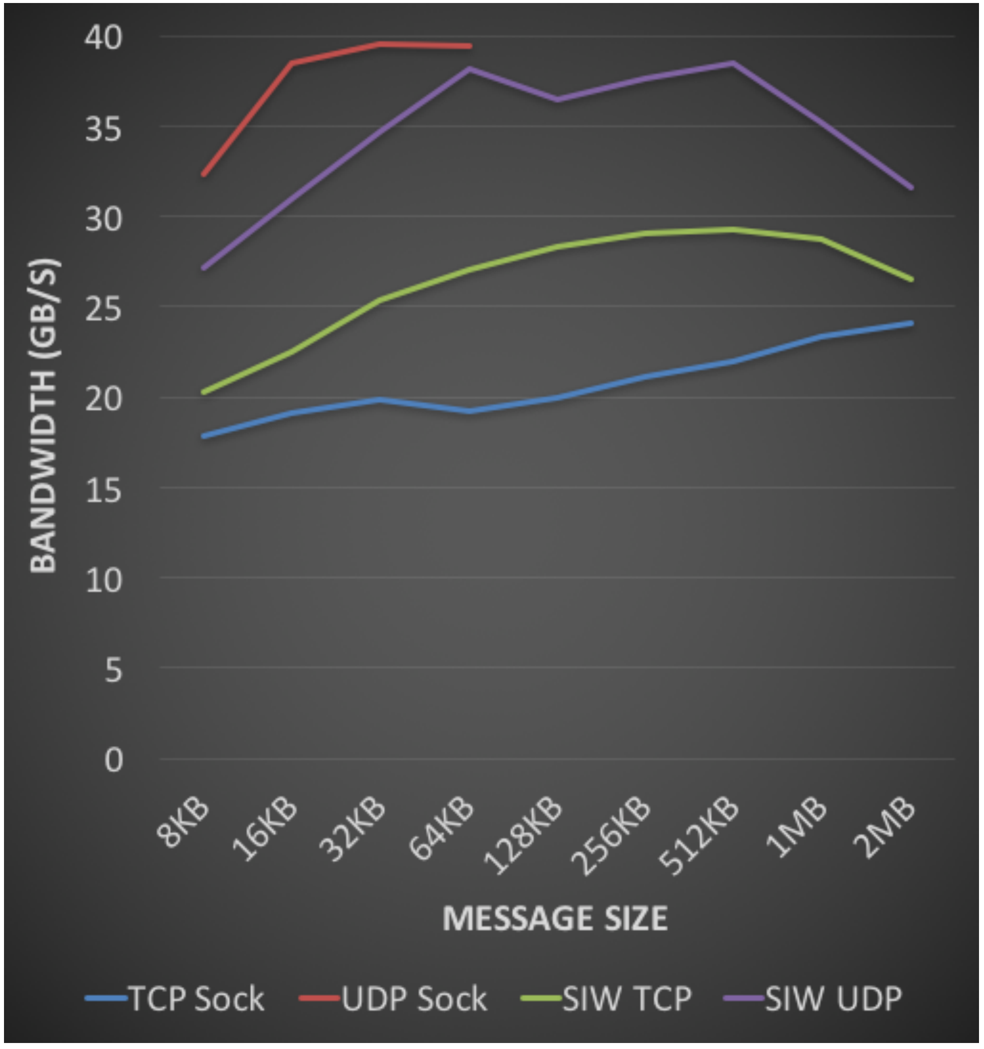
			Power (W)			
		BW (Gb/s)	Send cpu	Send dram	Recv cpu	Recv dram
TCP	Sock:	38.17	24.13	7.23	24.46	6.99
UDP	Sock:	39.55	24.14	9.59	23.35	7.09
SIWTCP	RDMA Read	35.66	25.6	4.8	24.64	5.36
SIWTCP	RDMA Write	23.83	22.65	6.18	11.48	5.53
SIWUDP	RDMA Read	39.17	21.55	6.64	21.7	5.3
SIWUDP	RDMA Write	38.33	26.58	8.03	13.34	6.98

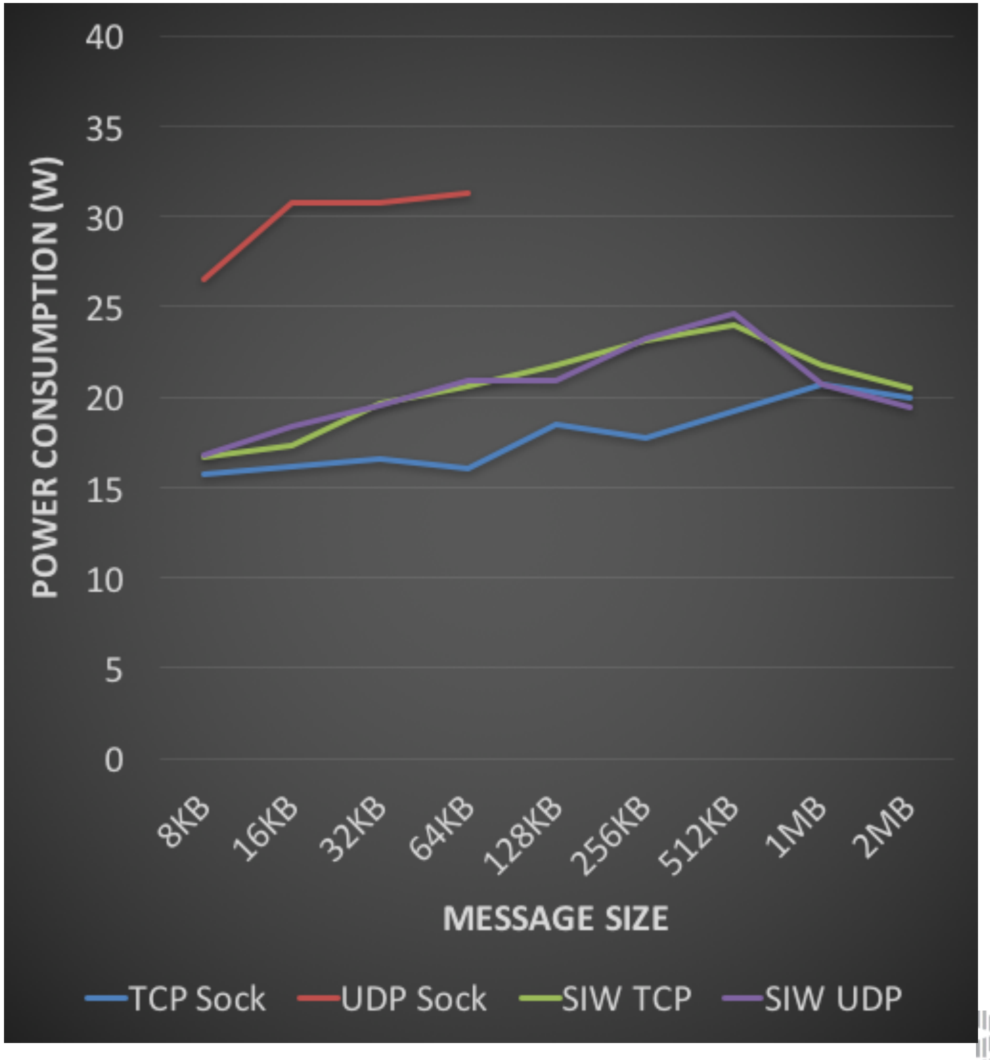
Power efficiency comparison

- We perform tests of all protocols and vary the message size
- Our power efficiency metric:

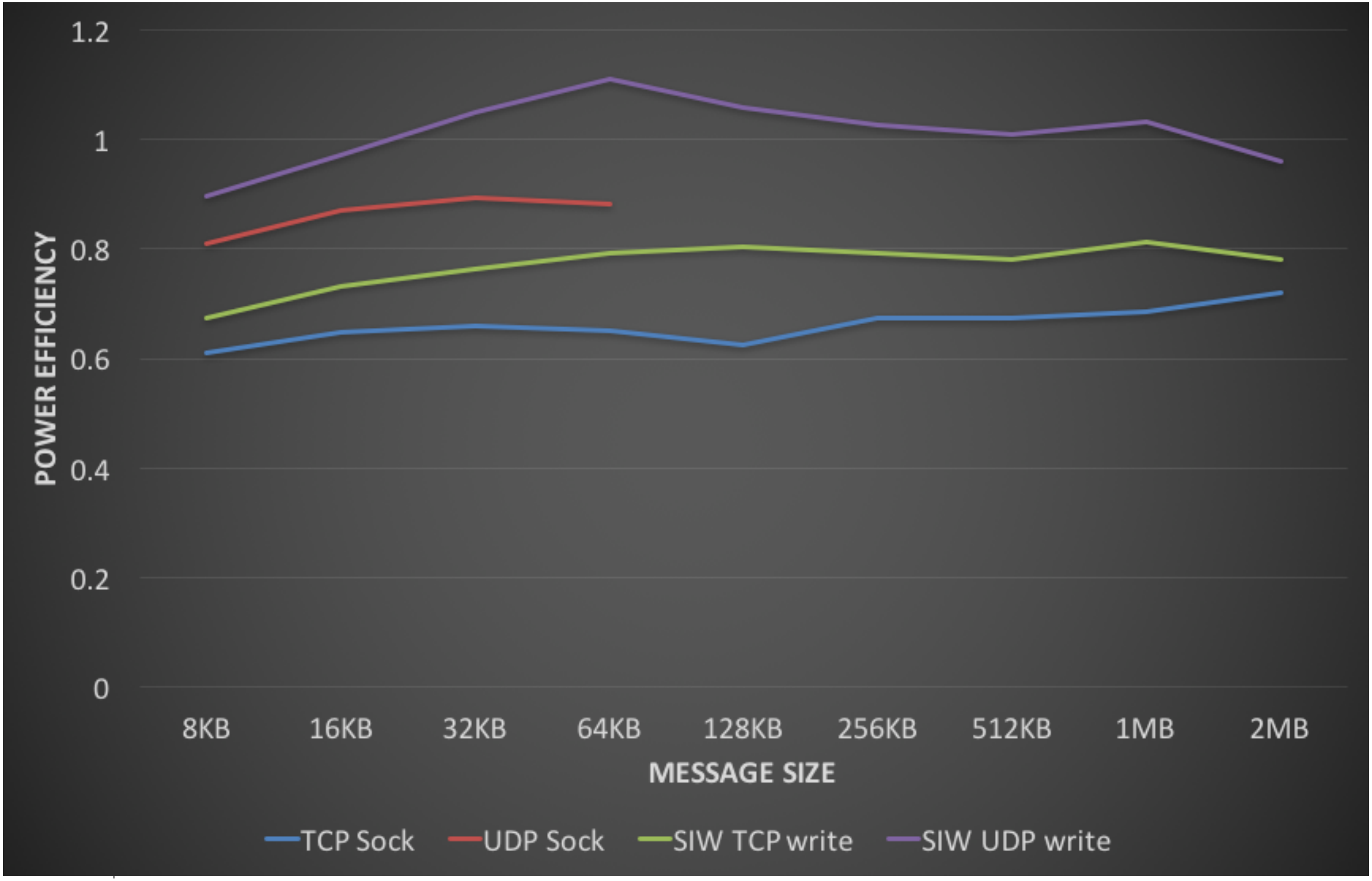
$$\frac{Bandwidth(Gb/s)}{PowerCons(Watt)}$$

Bandwidth and Power Consumption

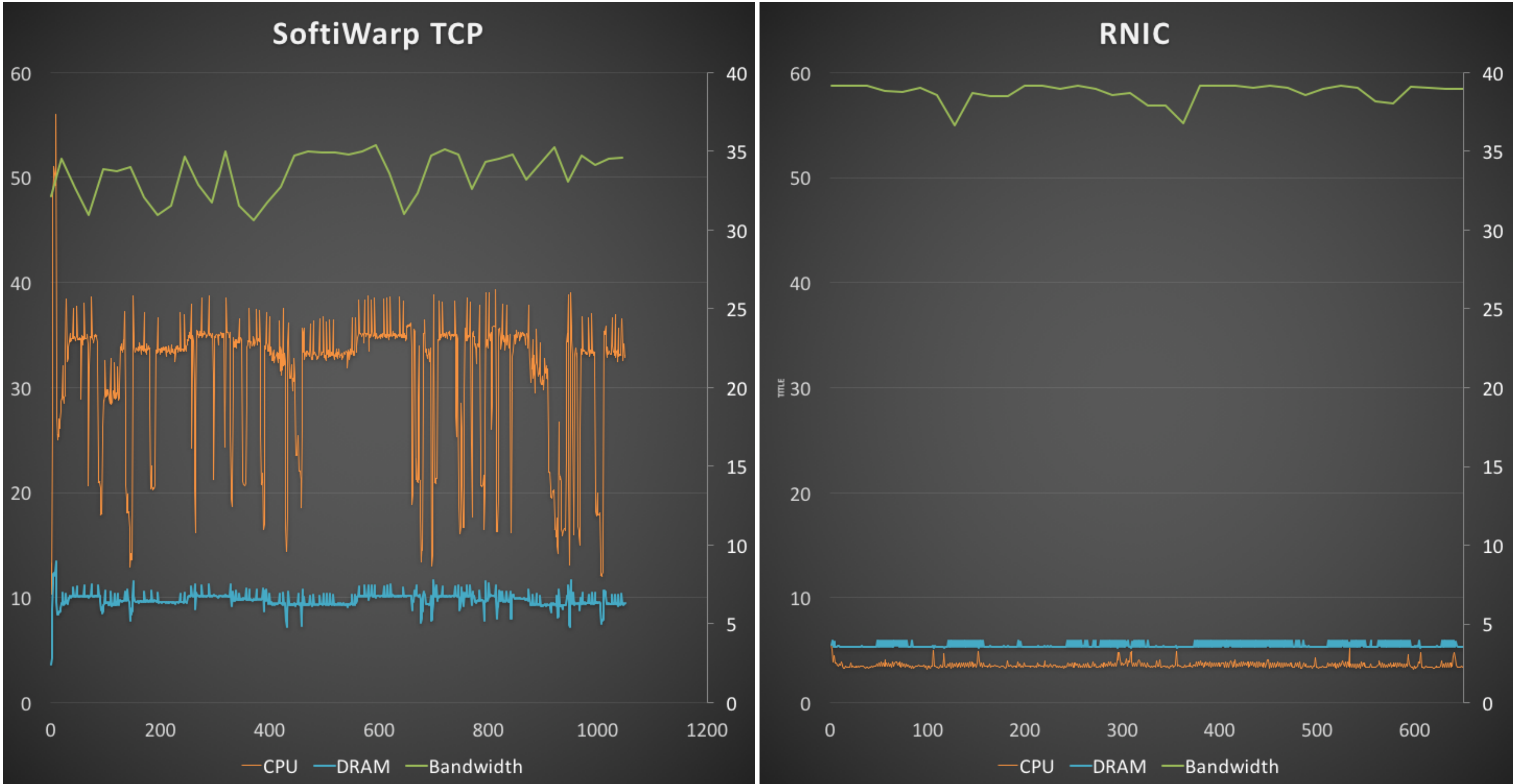




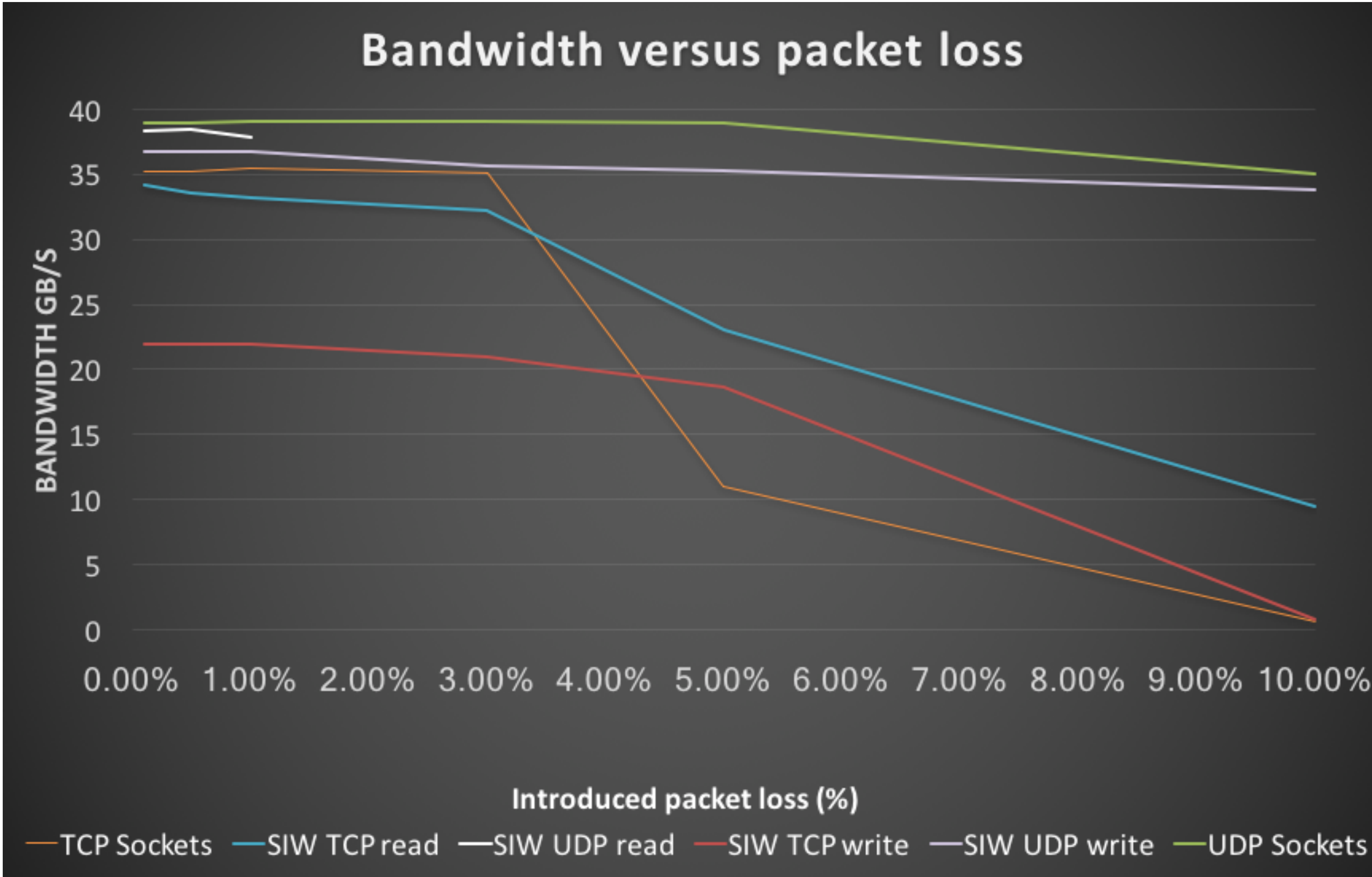
Calculated power efficiency (normalized power consumption)



Radio-astronomy data flow



Performance in case of packet loss


Bandwidth versus packet loss
BANDWIDTH GB/S
40
35
30
25
20
15
10
5
0
0.00%
1.00%
2.00%
3.00%
4.00%
5.00%
6.00%
7.00%
8.00%
9.00%
10.00%
Introduced packet loss (%)
TCP Sockets
SIW TCP read
SIW UDP read
SIW TCP write
SIW UDP write
UDP Sockets

Conclusions and future work

- An unreliable datagram-based iWARP protocol suits the requirements for radio-astronomy data transport service
- Software prototype looks promising
- Best power efficiency with hardware product (FPGA-based the most viable solution)
- Further work directions
  – Investigate the use of flash storage and big data frameworks (Spark, Hadoop)

Thank you!

http://www.dome-exascale.nl

lenkiewicz@nl.ibm.com

