# Power Efficiency for Software Algorithms running on Graphics Processors

Björn Johnsson

Per Ganestam

Michael Doggett

Tomas Akenine-Möller

LUND UNIVERSITY

# Overview



- Motivation
- Goal
- Project
  - Applications
  - Methodology
- Results
- Observations

# Motivation



- Energy efficency increasingly important
  - Phones, tablets, laptops, desktops...
- Hard area
  - Harder to analyze than regular performance
    - Need hardware and/or hardware support
  - Less intuitive / harder to predict

# Motivation



- Lots of papers looking at algorithms for power efficient hardware
- What can be done from the software side on current hardware?
- Learn more about energy and power

# Goal



- Say we want to optimize to lower energy usage
  - Do we have to measure power?
  - Or can we make an estimate based on rendering times alone?

# Project



- Two common graphics problems
- Implemented several different solutions solving those problems
- Measure their power usage and rendering time

# Applications



- Primary visibility and shading
  - –Forward rendering
  - –Forward rendering with pre-Z pass
  - –Deferred shading

# Applications

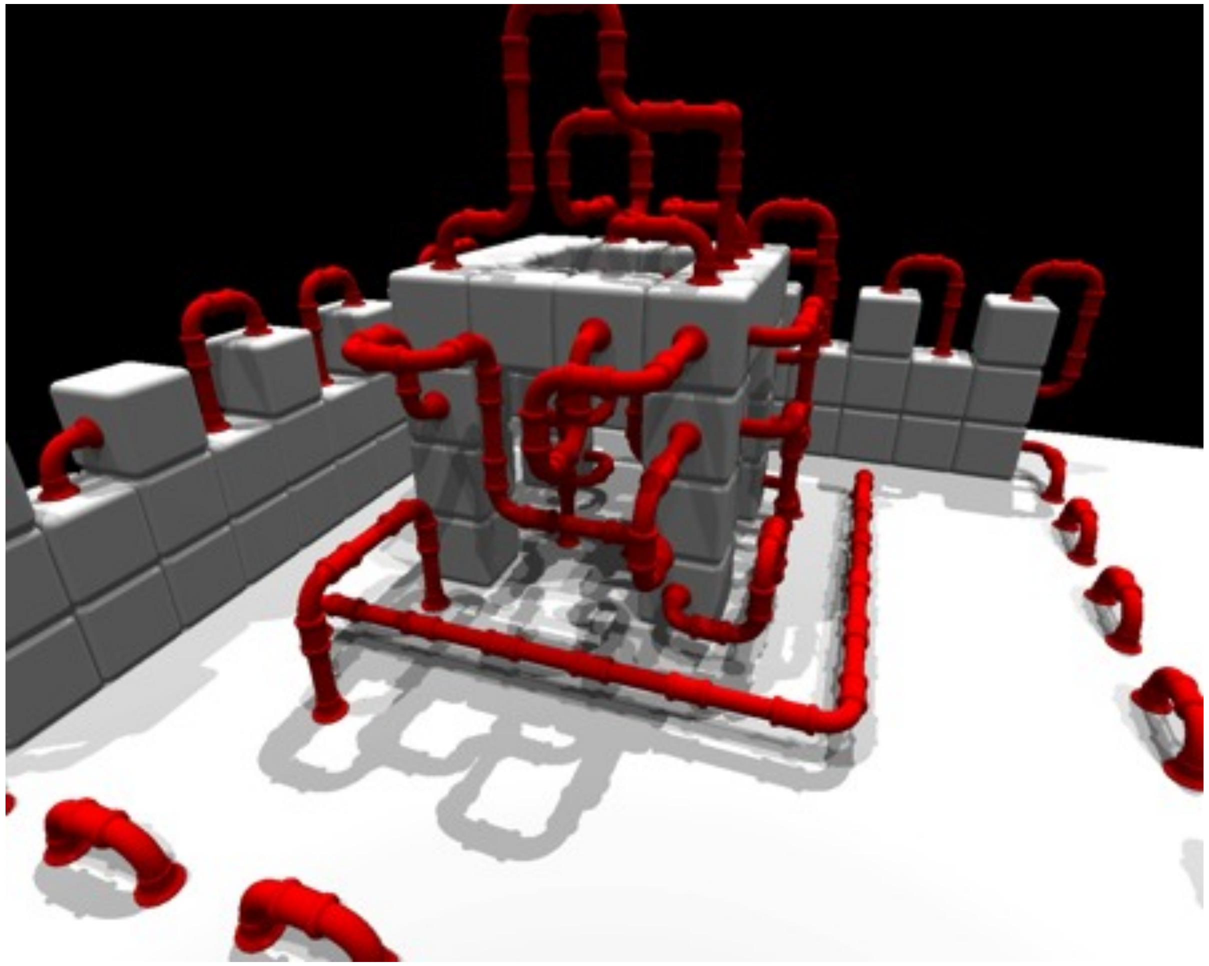


- Primary visibility and shading
  - –Forward rendering
  - –Forward rendering with pre-Z pass
  - –Deferred shading

# Applications



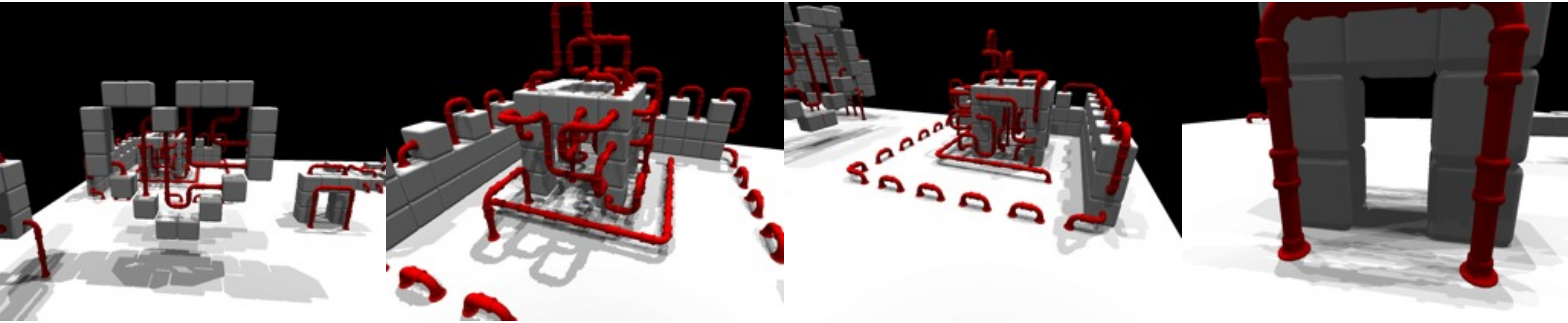
- Shadow algorithms
  - –Stencil shadow volumes
  - –Shadow mapping
  - –Variance shadow mapping



# Applications



- Shadow algorithms
  - –Stencil shadow volumes
  - –Shadow mapping
  - –Variance shadow mapping



# Applications

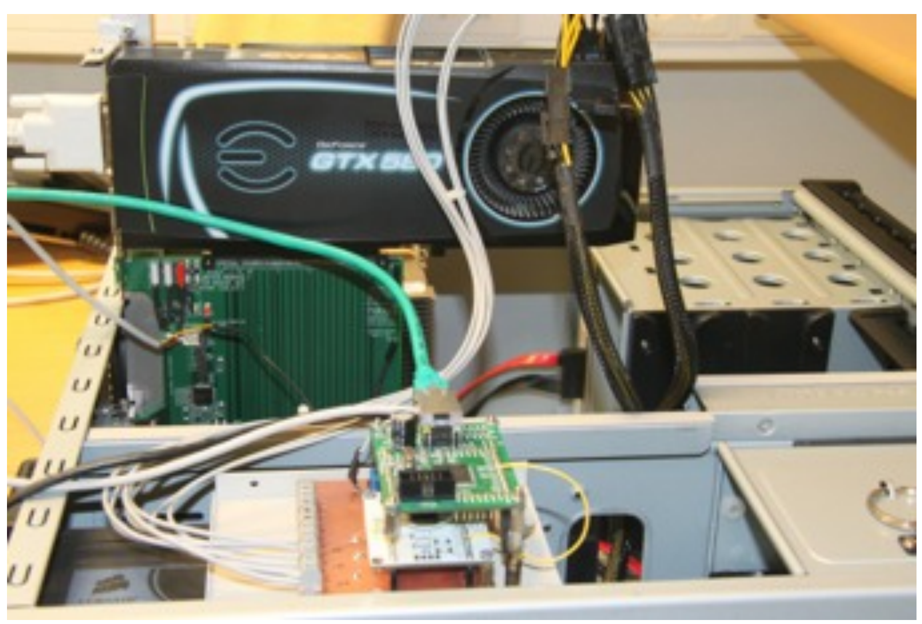


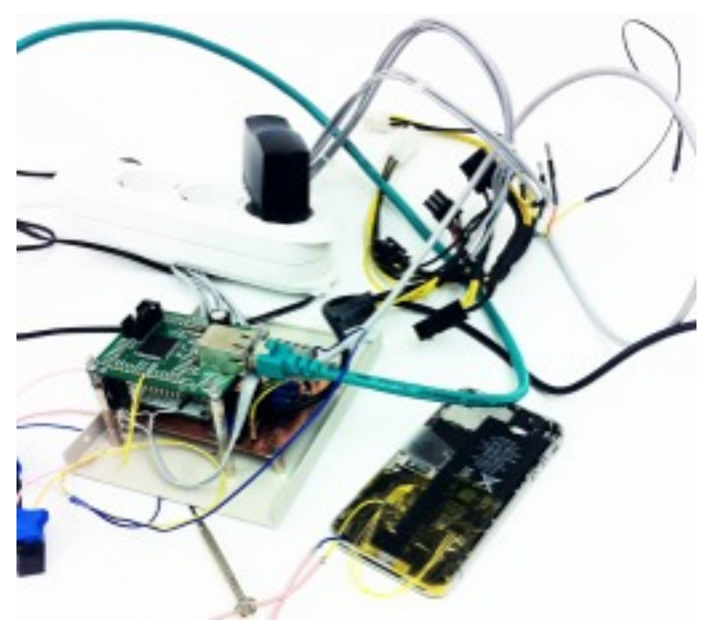
- OpenGL / OpenGL ES
- Full speed (no capping)
  - Widely different platforms gives different frame times (5ms to 2s)
- Timestamp at beginning and end of frame
  - Only integrate over actual rendering time
- Animated camera path (~2000 frames)

# Methodology



- We have built a measurement station
  - –Measure at 40kHz
  - –4 ACS710 Hall effect current sensors (<12A)
  - –2 shunt current sensors (<1A)
- Connected between GPU and power source
  - –Different places on different platforms

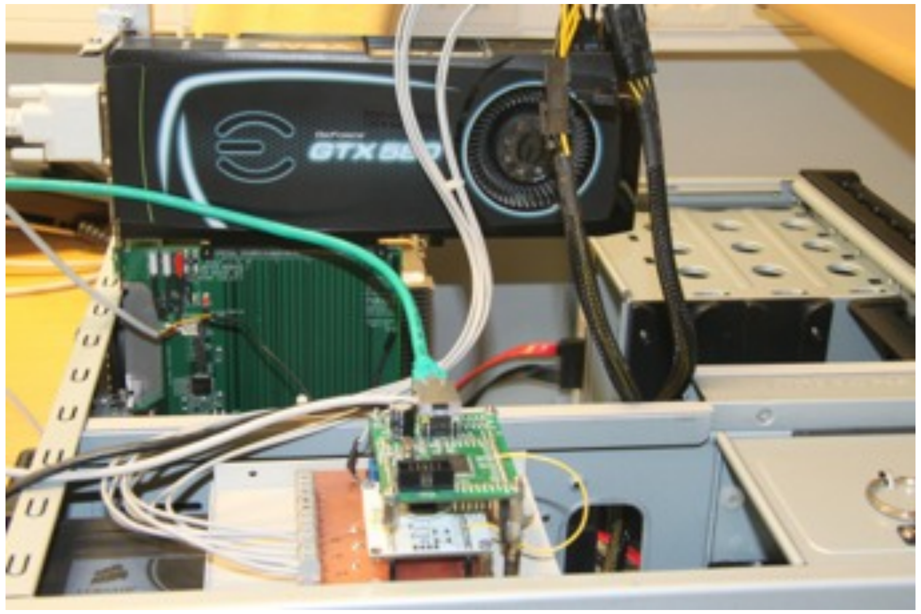




# Discrete graphics cards

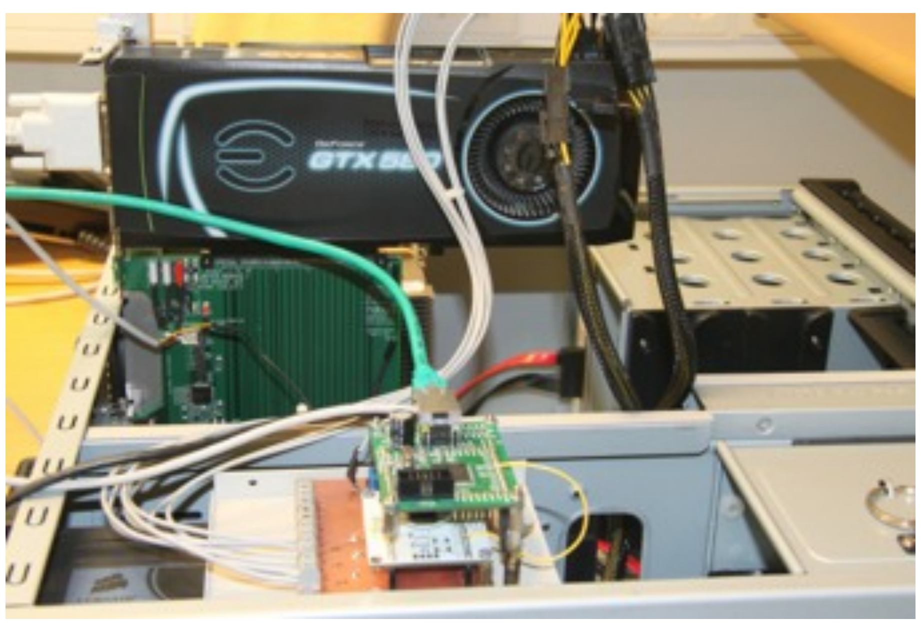


- Connected on the PCI-Express bus



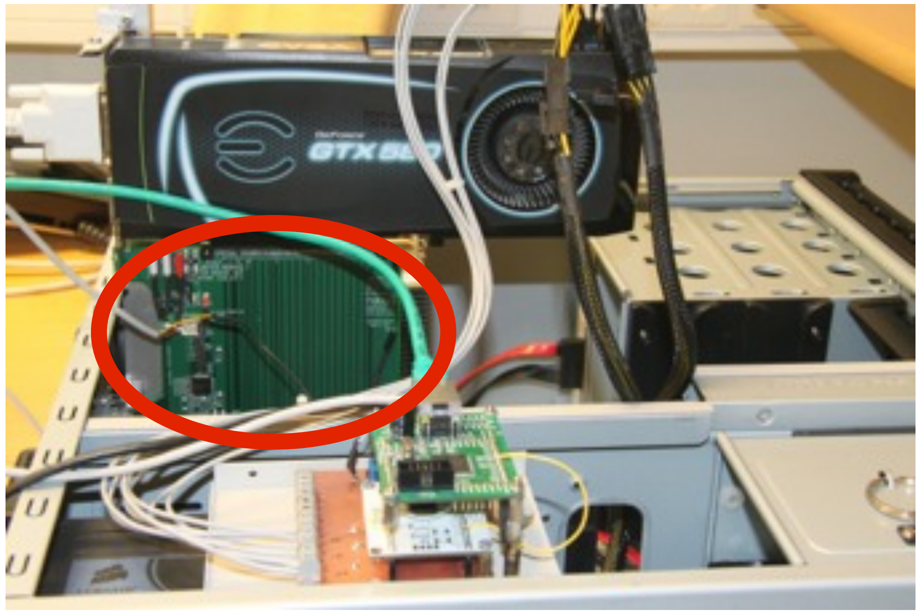
# Discrete graphics cards

- Connected on the PCI-Express bus
  - PCI-Express bus provides <=75W



# Discrete graphics cards

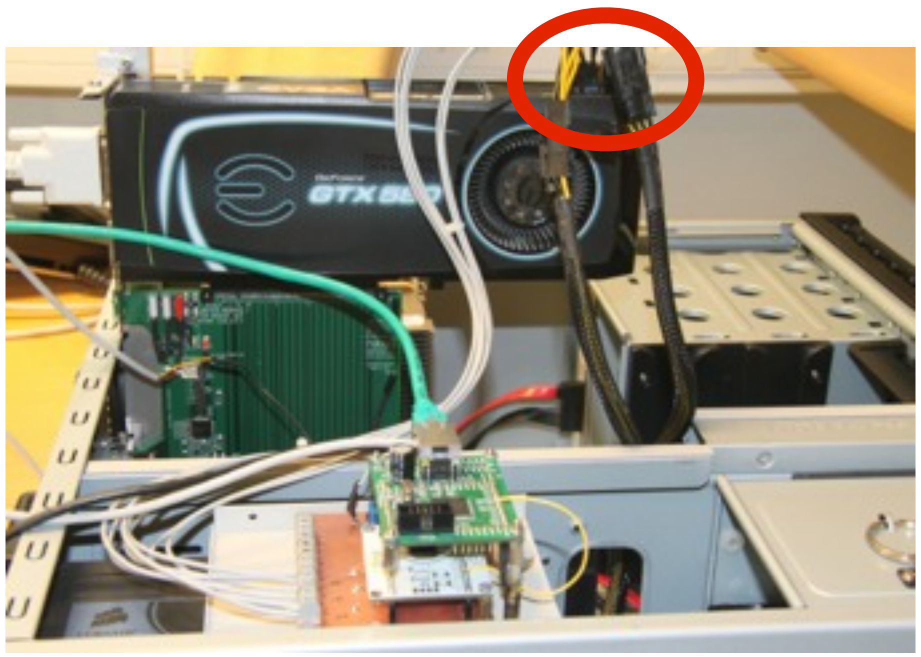
- Connected on the PCI-Express bus
  - PCI-Express bus provides <=75W
    - Measured through an Ultraview PCIeEXT-16HOT expander card



# Discrete graphics cards



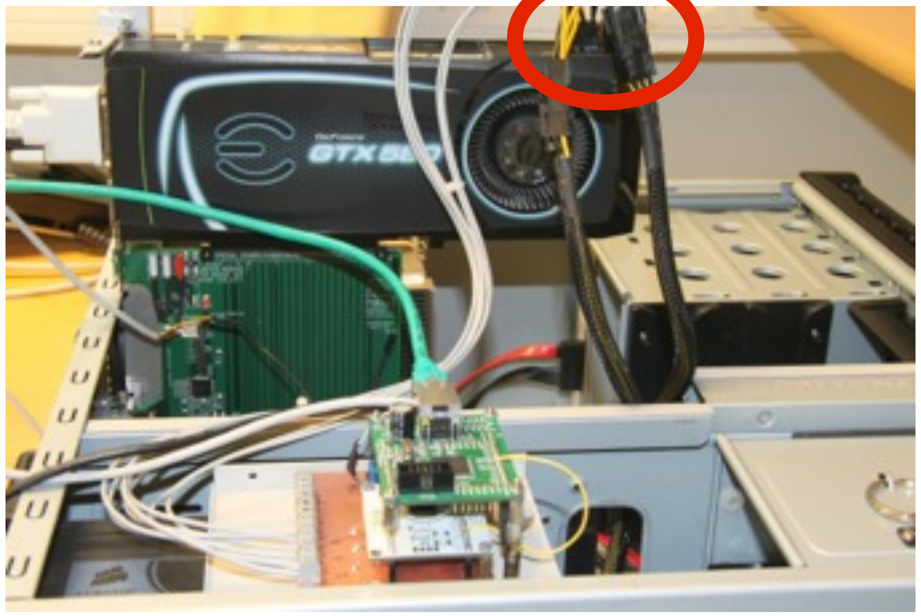
- Connected on the PCI-Express bus
  - PCI-Express bus provides <=75W
    - Measured through an Ultraview PCIeEXT-16HOT expander card
  - PCI-Express power connectors



# Discrete graphics cards

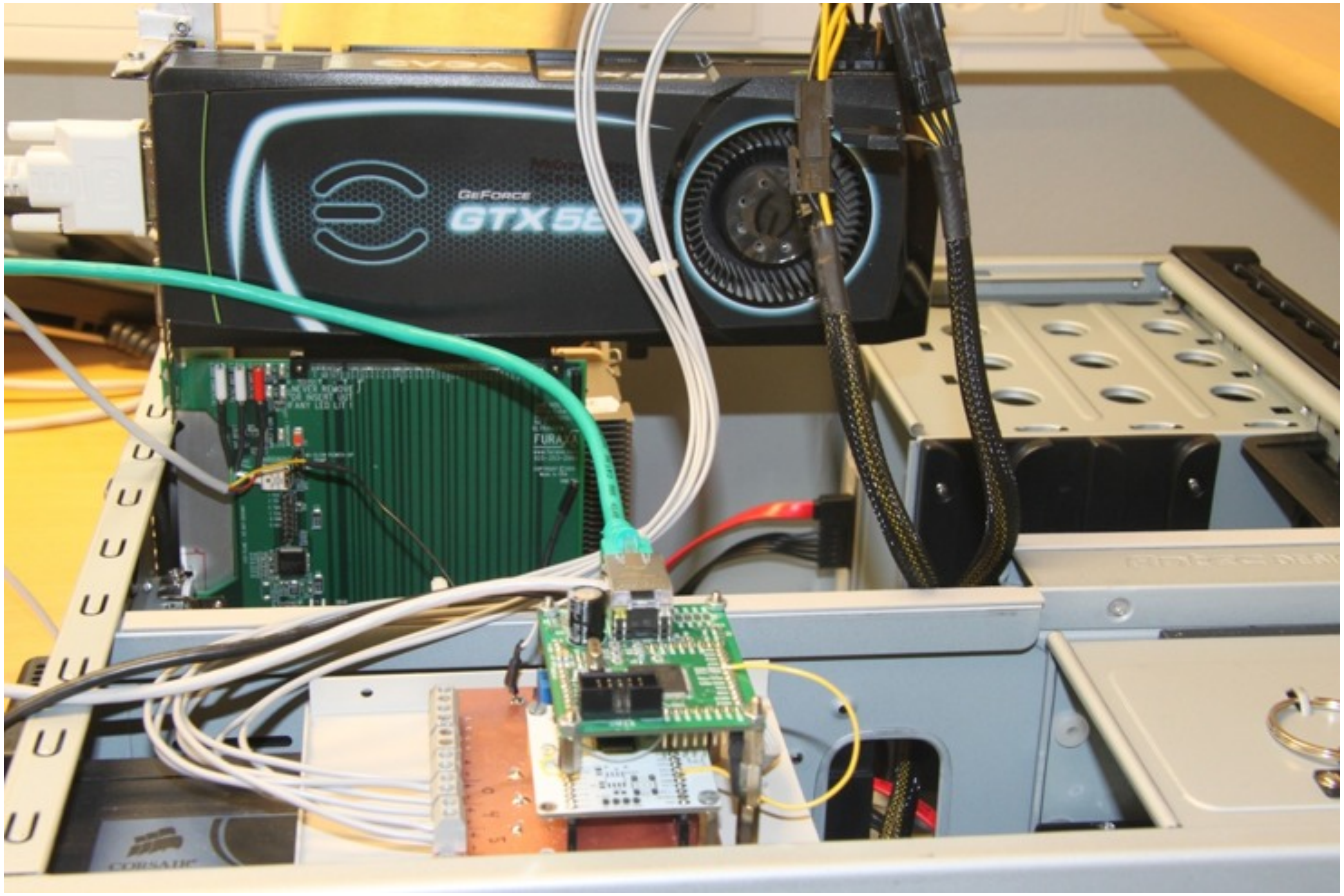


- Connected on the PCI-Express bus
  - PCI-Express bus provides 75W
    - Measured through an Ultraview PCIeEXT-16HOT expander card
  - PCI-Express power connectors
    - 8-pin provides <=150W
    - 6-pin provides <=75W



# Discrete graphics cards






LUND UNIVERSITY

# Discrete graphics cards



- AMD Radeon 7970 (28nm)
- NVIDIA GeForce GTX 580 (40nm)

# SoC #1



- Intel Sandy Bridge, HD3000 GPU (32nm)
  - Connected on motherboards 4-pin power connector
    - Provides power to CPU, GPU, and parts of the memory system
  - Two runs
    - Rendering pass
    - Idle pass, all gl* calls removed from code

# SoC #1

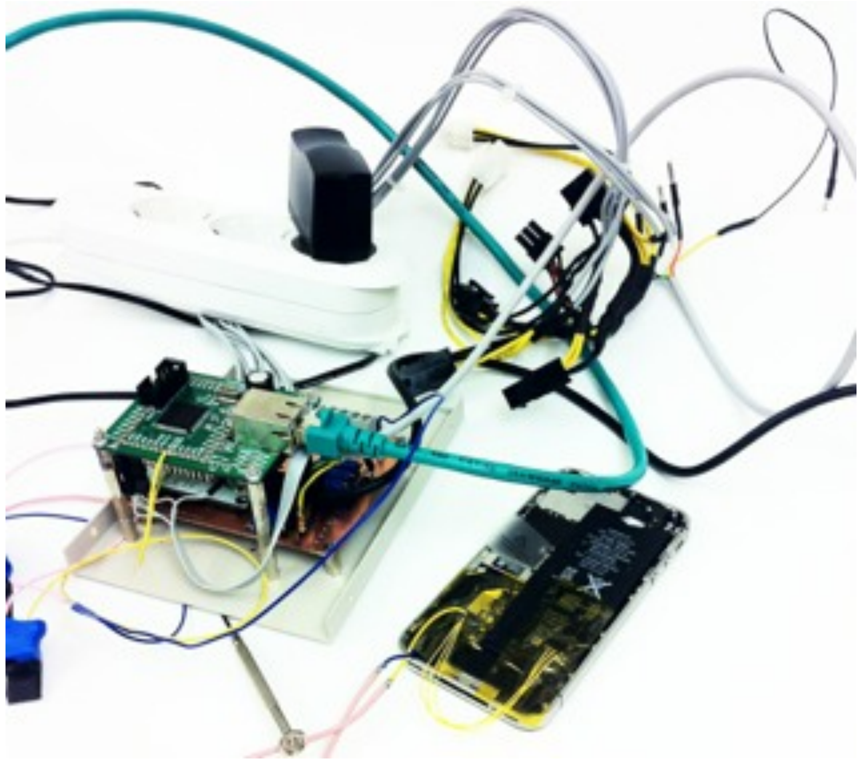


- Rendering pass - idle pass = ?
  - GPU power
  - Parts of the memory power
    - Memory bandwith generated from the graphics workload
    - Not including memory refresh power
  - CPU power for driver execution

# SoC #2



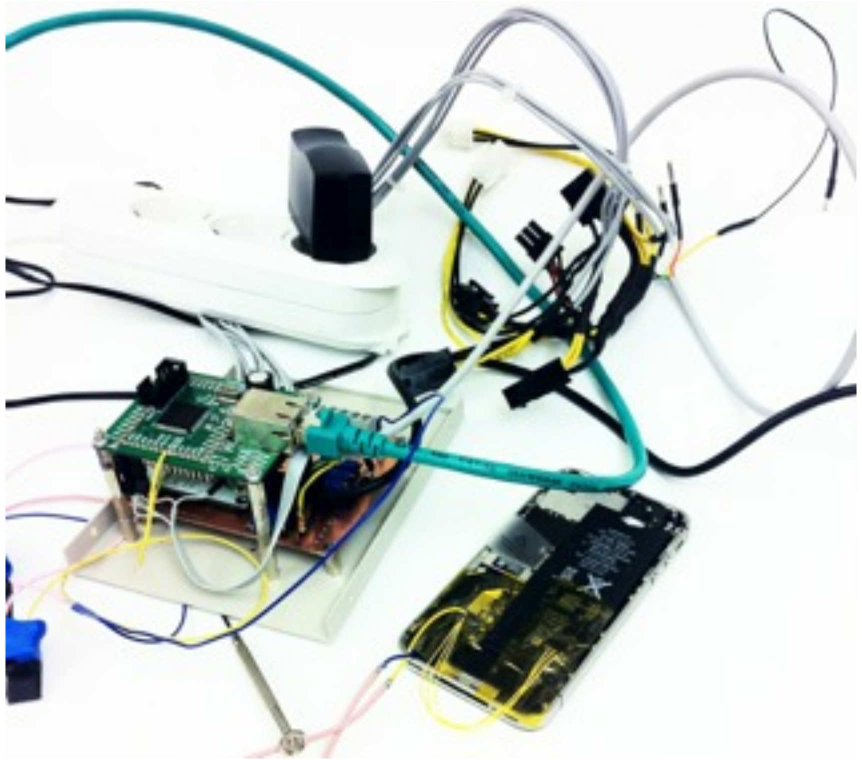
- iPhone 4S, PowerVR SGX543MP2 GPU (45nm)
  - Connected on battery connectors
    - Provides power to everything
  - Two runs
    - Rendering pass
    - Idle pass



# SoC #2

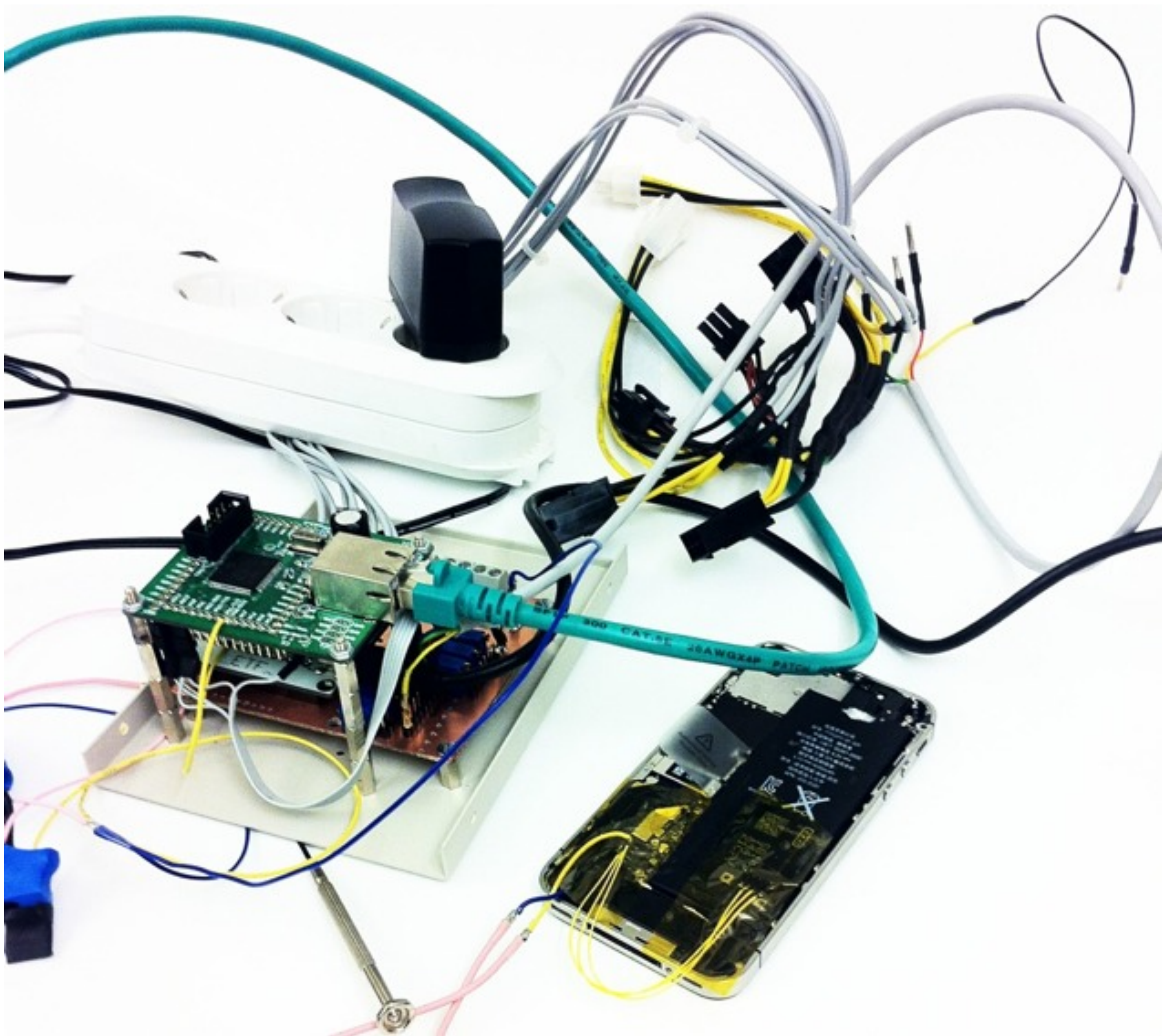


- Rendering pass - idle pass = ?
  – GPU power
  – Memory power
    - Only for memory bandwidth generated from the graphics workload
    - Not including memory refresh power
  – CPU power for driver execution



# SoC #2





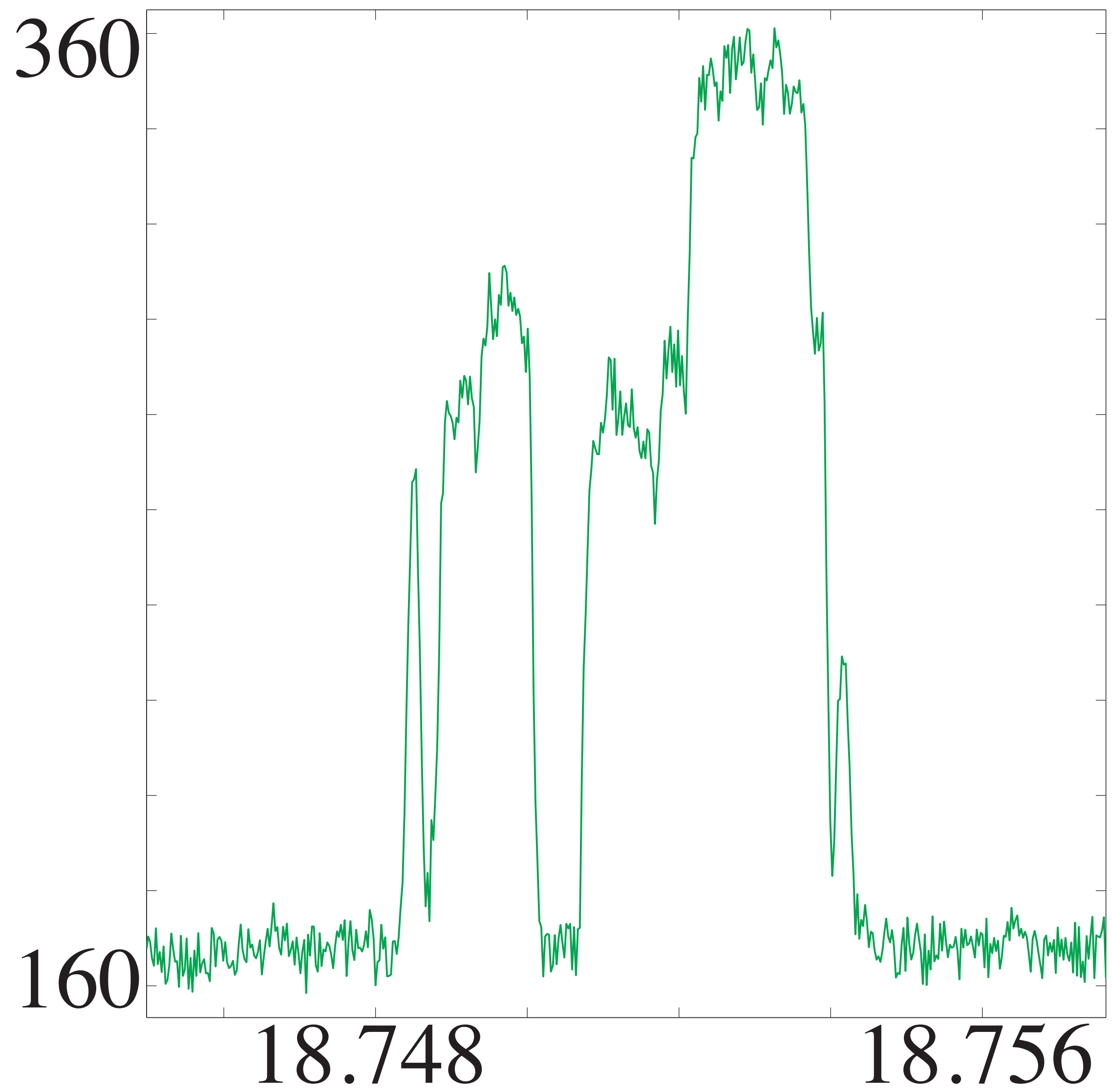
# Results



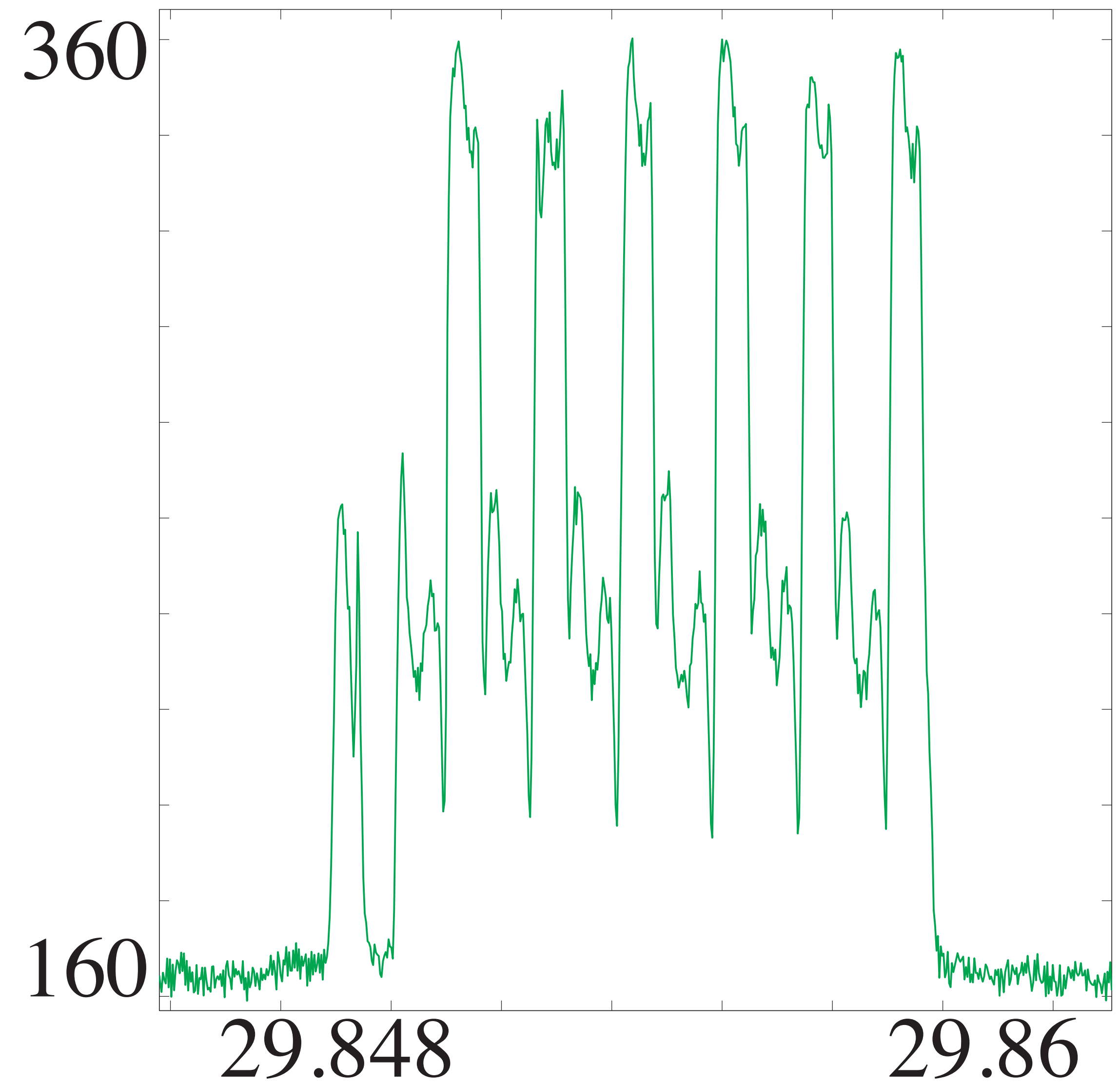
- What we measure:
  - High-frequency power data (40kHz)
  - Rendering time per frame

# GeForce GTX 580






Deferred shading



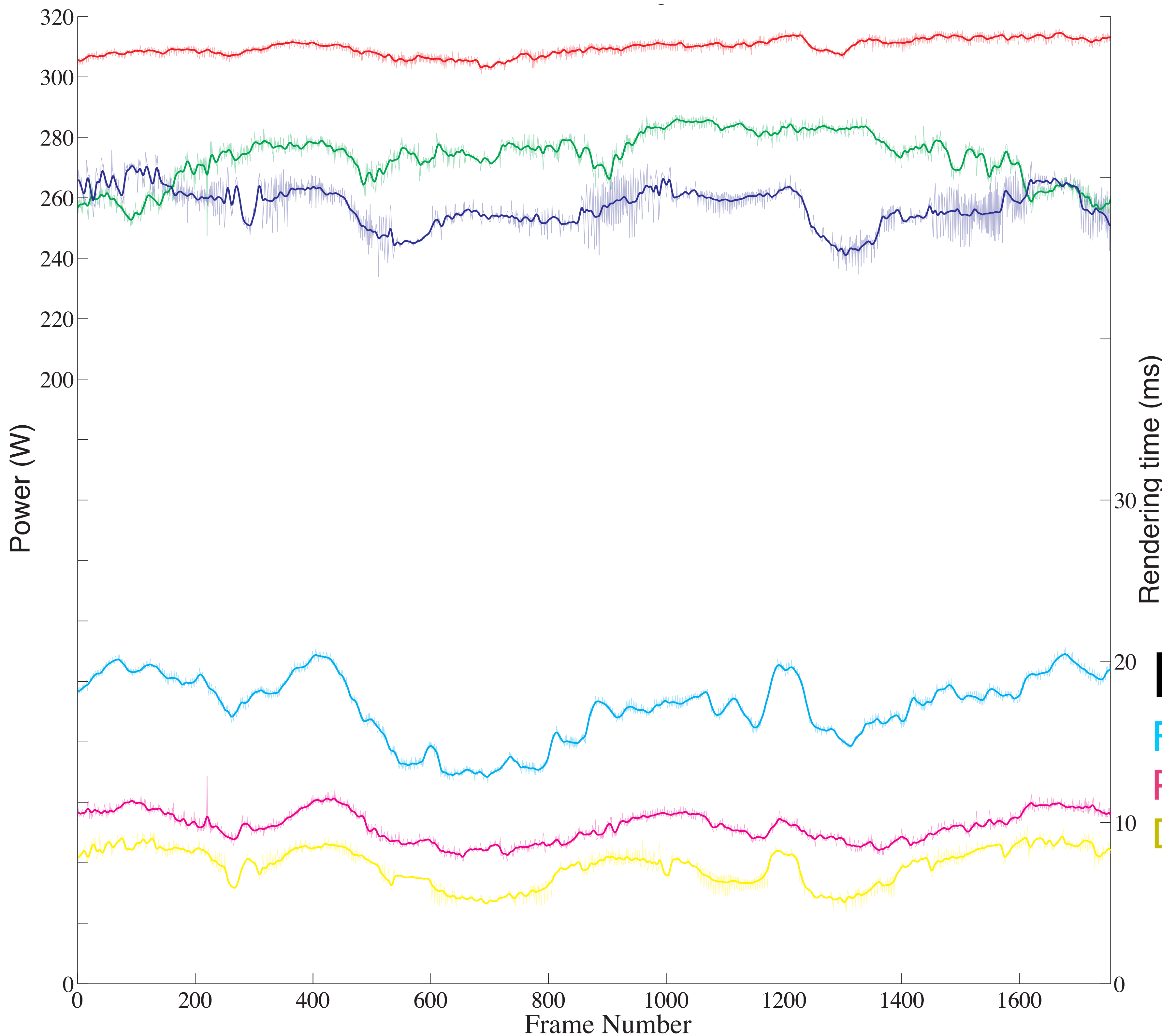

Variance shadow maps
6 light sources

LUND UNIVERSITY

# GeForce GTX 580: Primary rendering



**Power**
Forward
Pre-Z
Deferred



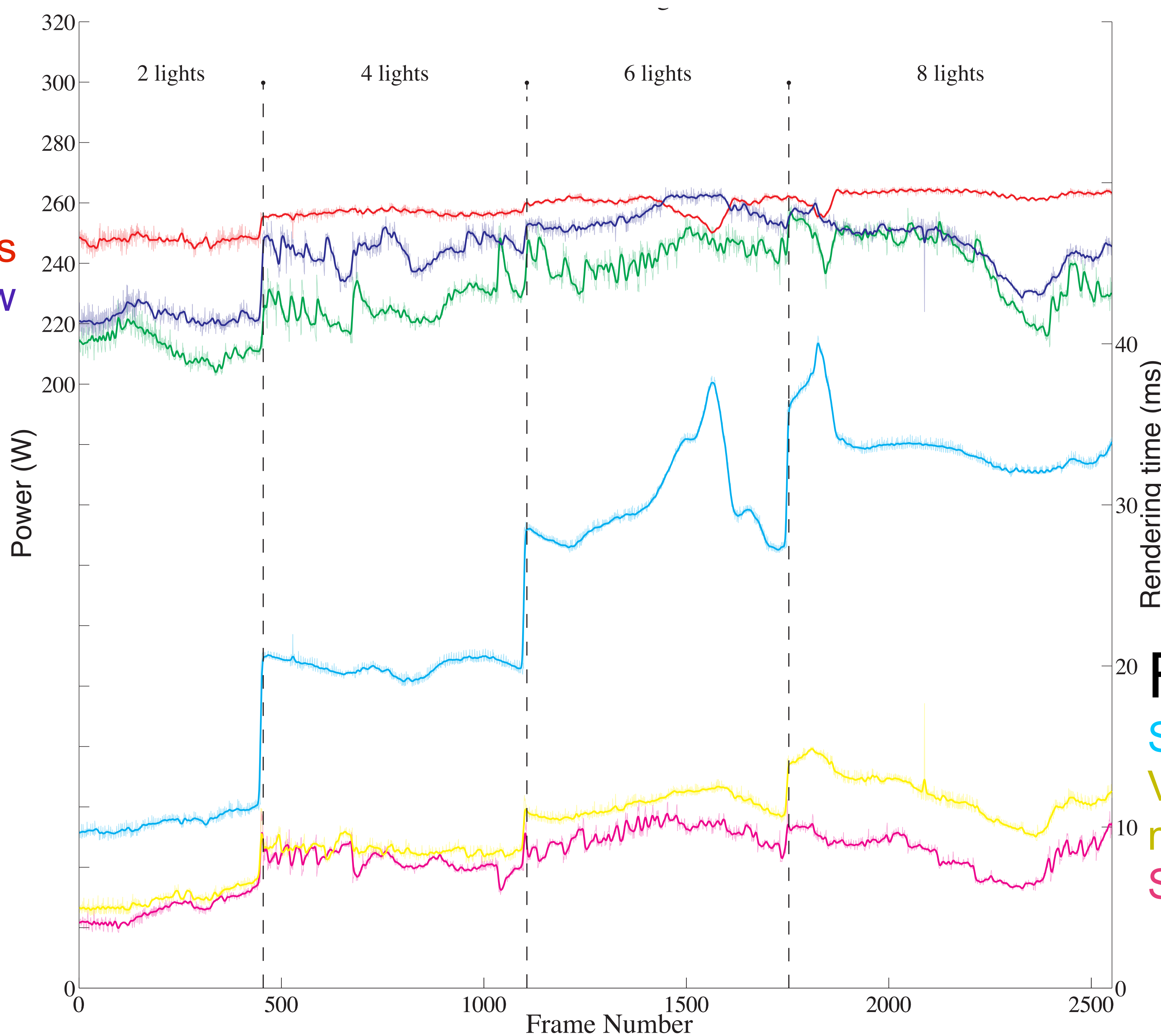

**Frametime**
Forward
Pre-Z
Deferred

LUND UNIVERSITY

# GeForce GTX 580: Shadows



Power
Shadow volumes
Variance shadow maps
Shadow maps




Frametime
Shadow volumes
Variance shadow maps
Shadow maps

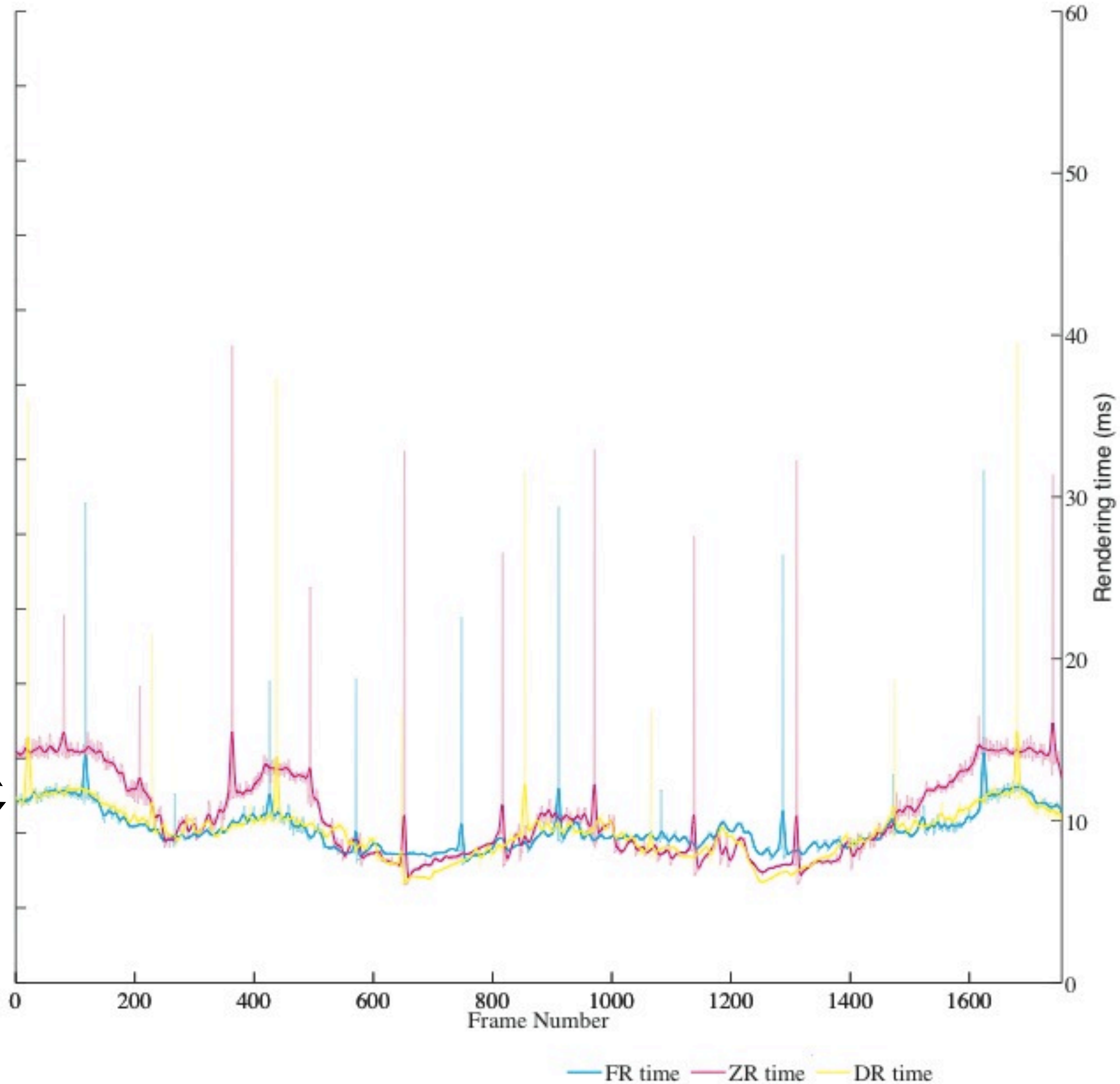
LUND UNIVERSITY

# AMD Radeon 7970: Primary rendering



Forward

Deferred

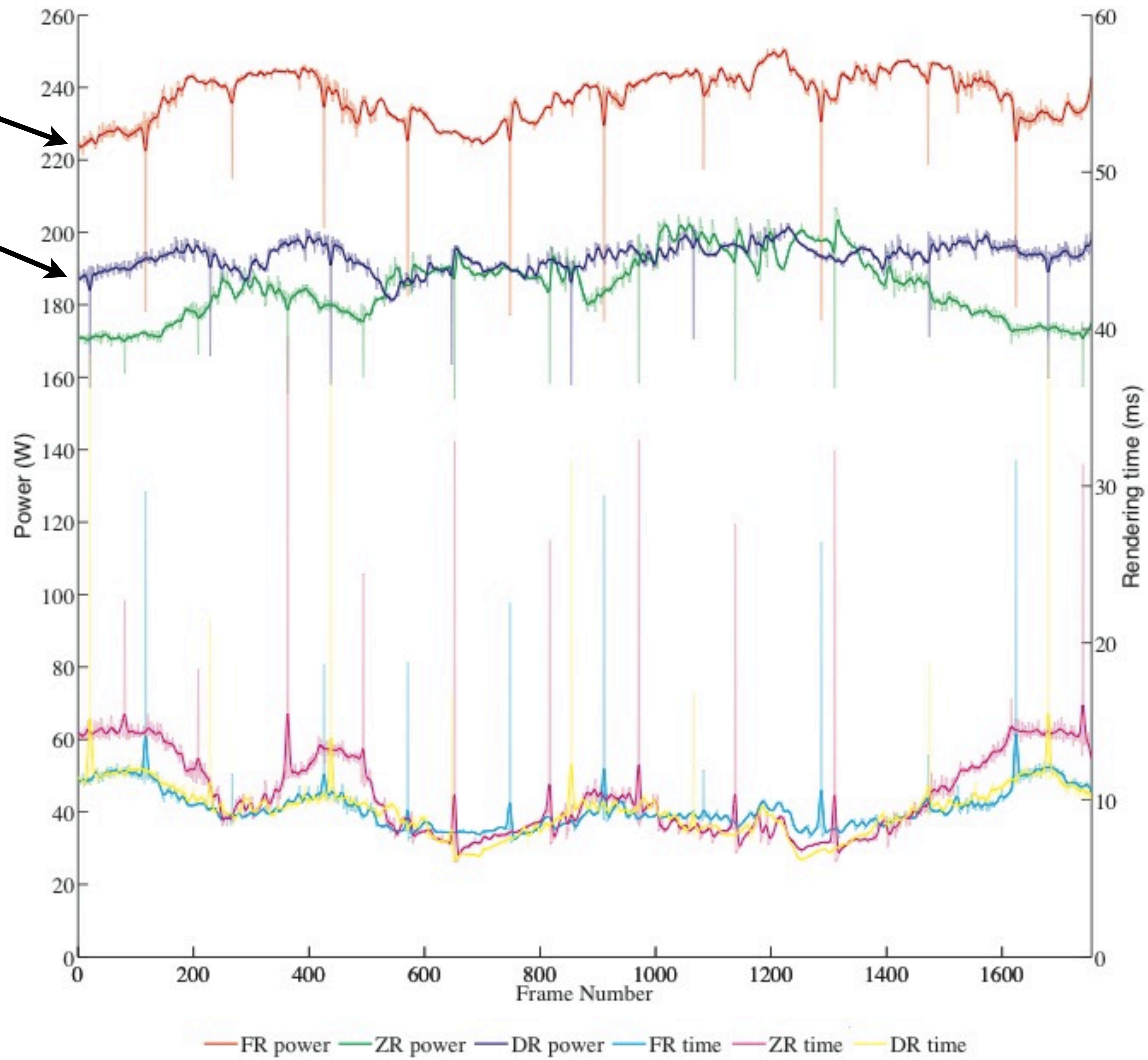


# AMD Radeon 7970: Primary rendering



Forward

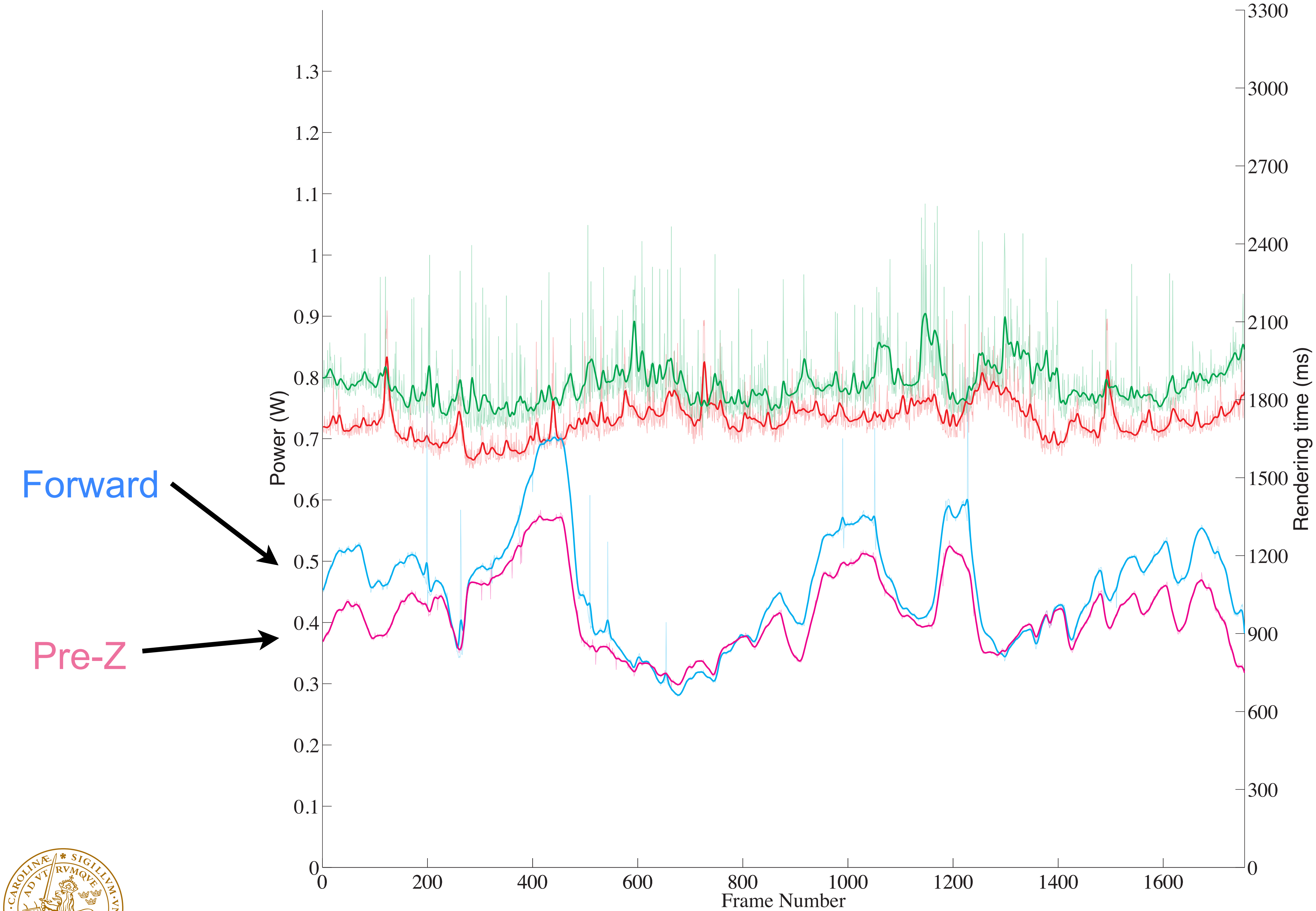
Deferred




LUND UNIVERSITY

# iPhone 4S




Forward
Pre-Z
Power (W)
Rendering time (ms)
Frame Number
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2
1.3
0
300
600
900
1200
1500
1800
2100
2400
2700
3000
3300
0
200
400
600
800
1000
1200
1400
1600

# iPhone 4S

- Higher energy than expected
  - –Mostly due to long rendering times
  - –Probably pushing sort-middle to flush buffers

Forward

Pre-Z

# Results



- What numbers are we interested in?

# Results



- What other numbers are we interested in?
  - Power

# Results



- What other numbers are we interested in?
  - –Power
  - –Energy

# Results



- What other numbers are we interested in?
  - Power
  - Energy
    - More interesting for battery lifetime

# Results



- What other numbers are we interested in?
  - Power
  - Energy
    - More interesting for battery lifetime
    - But what metric should we use?

# Metric



nJ/pixel

# Metric



nJ/pixel

- Largely resolution independent

# Metric



nJ/pixel

- Largely resolution independent
- Easy to divide frame into segments

# Metric



nJ/pixel

- Largely resolution independent
- Easy to divide frame into segments
- Easy to calculate fillrate for a given TDP

# Energy, nJ/pixel



| | Primary visibility | | |
|---|---|---|---|
| | FR | ZR | DR |
| GTX 580 | 1,443 | 722 | 511 |
| Radeon 7970 | 607 | 512 | 489 |
| Sandy Bridge | 872 | 314 | 280 |
| iPhone 4S | 2,234 | 2,015 | - |

# Energy, nJ/pixel

| | Shadows | | |
|---|---|---|---|
| | SV | SM | VSM |
| GTX 580 | 1,325 | 447 | 532 |
| Radeon 7970 | 953 | 469 | 804 |
| Sandy Bridge | 1,317 | 311 | 511 |
| iPhone 4S | - | 461 | - |

# Observations



- Not possible to estimate power / energy only from frame times
- Surprisingly, pre-Z proved useful on a tiled, deferred, sort-middle architecture for our application
  - Pushing too much geometry?
- Still similar energy per pixel, despite rendering times that differ by an order of magnitude or more
- nJ/pixel

Thanks for listening

Any questions?