Seminar: Scheduling – Modern Problems in a Seemingly Solved Discipline

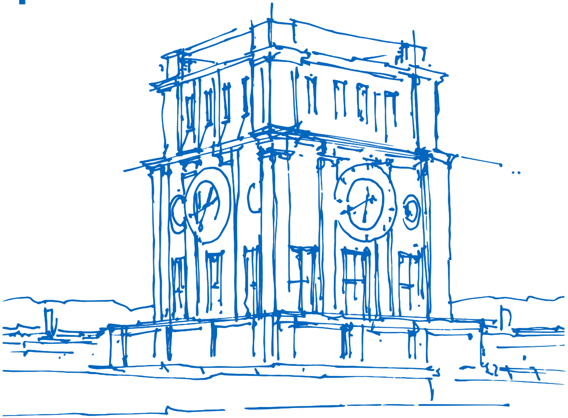
Kick-off

Sommersemester 2022

Dr. Matthias Maiterth
Dr. Eishi Arima
Dr. Isaías Comprés

Computer Architecture & Parallel Systems
Fakultät für Informatik
Technische Universität München

Mai 2nd, 2022



Kick-off – Outline

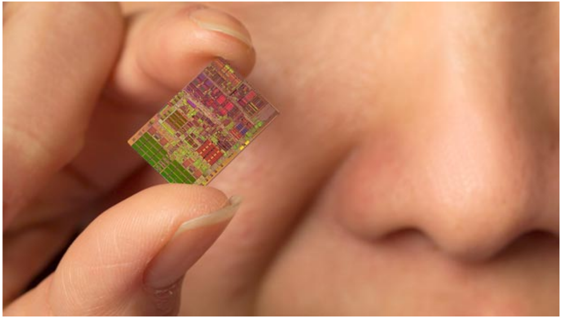
- Why scheduling is a hot topic (and continues to be!)
- Introducing the organizers
- Assignments overview
- Semester outline
- Topics presentation and selection
- Next steps

"And you have to realize that there are not very many things that have aged as well as the scheduler. Which is just another proof that scheduling is easy."

Linus Torvalds, 2001. [1] [1]

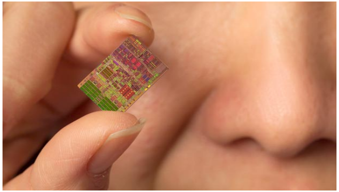
[1] [1]: http://tech-insider.org/linux/research/2001/1215.htm

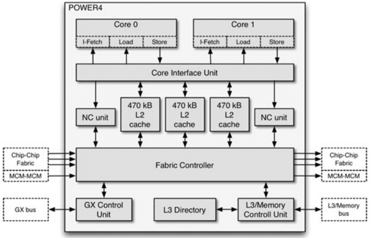
Also 2001:



Why scheduling is a hot topic
(And continues to be a hot topic!)

- Process scheduling: from single-core to multi-core

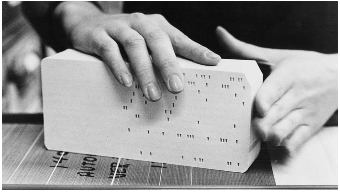




Source: https://www.ibm.com/ibm/history/ibm100/us/en/icons/power4/

Why scheduling is a hot topic
(And continues to be a hot topic!)

- Process scheduling: from single-core to multi-core
- Batch processing: from Punch-Card programmed computers to modern Supercomputers

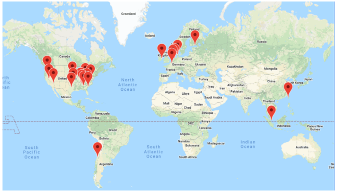


left: https://www.ibm.com/ibm/history/ibm100/us/en/icons/punchcard/breakthroughs/
right: https://doku.lrz.de/display/PUBLIC/SuperMUC-NG

Computer Architecture & Parallel Systems
Fakultät für Informatik
Technische Universität München

Why scheduling is a hot topic
(And continues to be a hot topic!)

- Process scheduling: from single-core to multi-core
- Batch processing: from Punch-Card programmed computers to modern Supercomputers
- Resource Management: from Multi-User to Geographically Distributed Cluster Management



left: https://www.ibm.com/ibm/history/ibm100/us/en/icons/system360/
right: https://www.google.com/about/datacenters/locations/

Introducing the organizers

- Dr. Eishi Arima
  - Research:
    - Computer architecture and parallel systems
    - Power-aware computing
    - Process and resource management in HPC systems
    - Heterogeneous systems
    - Memory and storage systems
    - Performance modeling, analysis, and optimization
    - Hardware/software codesign
    - Processor microarchitectures
  - Project: REGALE

Introducing the organizers

- Dr. Isaías Comprés
  - Research:
    - Supercomputing
    - Message Passing Interface (MPI) Standard
    - Distributed Memory Programming Models
    - Process Management Interface Exascale (PMIx) Standard
    - Process and Resource Management
    - Batch Scheduling
    - Distributed Memory Tools
    - Online Automatic Tuning
  - Project: DEEP-SEA

Introducing the organizers

- Dr. Matthias Maiterth
  - Research:
    - Energy Efficiency
    - High Performance Computing
    - Resource Management
    - Computer Architecture
  - Teaching:
    - Introduction to Computer Architecture (ERA)

Assignments overview

1. Presentation:
   - 20 minute (Bachelors)
   - 25 minute (Masters)

   2 presentations per week

2. Seminar paper:
   - 4 page paper (Bachelors)
   - 6 page paper (Masters)

   double column, IEEEtran

Goal of this seminar:

- Give
  - an introduction
  - the necessary background
  - current challenges
  - discussion and
  - an outlook into the future

  of your topic.
- Gain knowledge in the research areas of scheduling by actively participating

Semester outline

Date	Session	Paper – due dates
02.05.2022	Introduction and topic assignment	
09.05.2022	Guidelines for scientific writing and presentations	
16.05.2022	Free / buffer	
23.05.2022	Student presentations (1 & 2)	
30.05.2022	Student presentations (3 & 4)	
06.06.2022	Whit Monday / Holiday	
13.06.2022	Student presentations (5 & 6)	Initial seminar-paper submission
20.06.2022	Student presentations (7 & 8)	
27.06.2022	Student presentations (9 & 10)	Review Period
04.07.2022	Student presentations (11 & 12)	
11.07.2022	Student presentations (13 & 14)	
18.07.2022	Student presentations (15 & buffer)	Camera Ready Version
25.07.2022	Summary	(Seminar proceedings)

Kick-off – Outline

- Why scheduling is a hot topic (and continues to be!)
- Introducing the organizers
- Assignments overview
- Semester outline
- **Topics presentation and selection**
- Next steps

Topics

1. Metrics and Fundamentals of Scheduling
2. Threads and Tasks Scheduling
3. GPU Scheduling (Warps and Compute Units)
4. Process Scheduling (OS)
5. Modern Batch Scheduling by example of FLUX
6. Clouds and Container Scheduling
7. Large-Scale Cluster management by example of Borg
8. Scheduling for Heterogeneous Systems
9. Workflows and scheduling
10. Co-Scheduling
11. IO-Aware Scheduling
12. Malleability
13. Power-Aware Scheduling
14. ML-Based Scheduling
15. Scheduling for Quantum Systems

1. Metrics and Fundamentals of Scheduling

- Feitelson, Rudolph and Schwiegelshohn,
  "Parallel Job Scheduling — A Status Report",
  JSSPP 2004 (https://link.springer.com/chapter/10.1007/11407522_1)
- Feitelson, et.al.,
  "Theory and Practice in Parallel Job Scheduling",
  JSSPP 1997 (https://www.cs.huji.ac.il/w~feit/parsched/jsspp97/p-97-1.pdf)
- Ngubiri and van Vliet,
  "Characteristics of Fairness Metrics and their Effect on Perceived Scheduler Effectiveness",
  IJCA 2010 (https://www.tandfonline.com/doi/abs/10.1080/1206212X.2010.11441974)
- Gómez-Martín, Vega-Rodríguez and González-Sánchez,
  "Fattened backfilling: An improved strategy for job scheduling in parallel systems", JPDC 2016 (https://www.sciencedirect.com/science/article/abs/pii/S0743731516300788)

2. Threads and Tasks Scheduling

- Iwasaki, et.al.,
  "BOLT: Optimizing OpenMP Parallel Regions with User-Level Threads",
  PACT'19 (`https://pavanbalaji.github.io/pubs/2019/pact/pact19.bolt.pdf`)
- Bak, et.al.
  "Task-graph scheduling extensions for efficient synchronization and communication",
  ISC'21 (`https://dl.acm.org/doi/abs/10.1145/3447818.3461616?sid=SCITRUS`)
- Korndörfer, et.al.,
  "LB4OMP: A Dynamic Load Balancing Library for Multithreaded Applications",
  TPDS'22 (`https://ieeexplore.ieee.org/document/9524500`)

3. GPU Scheduling (Warps and Compute Units)

- Narasiman, et.al,
  "Improving GPU Performance via Large Warps and Two-Level Warp Scheduling",
  MICRO, 2011 (https://users.ece.cmu.edu/~omutlu/pub/large-gpu-warps_micro11.pdf)
- Brunie, Collange and Diamos,
  "Simultaneous Branch and Warp Interweaving for Sustained GPU Performance",
  ISCA 2012 (https://hal-ens-lyon.archives-ouvertes.fr/ensl-00649650v2/document)
- Ausavarungnirun, et.al.,
  "Exploiting Inter-Warp Heterogeneity to Improve GPGPU Performance",
  PACT'15 (https://dl.acm.org/doi/10.1109/PACT.2015.38)

4. Process Scheduling (OS)

- Lepers, Quéma and Fedorova,
  "Thread and memory placement on NUMA systems: asymmetry matters",
  USENIX ATC 2015 (https://www.usenix.org/system/files/conference/atc15/atc15-paper-lepers.pdf)
- Nesterenko, Qi and Rao,
  "Improving Resource Utilization through Demand Aware Process Scheduling",
  ICPP'18 (https://dl.acm.org/doi/10.1145/3225058.3225132)
- Yu, et.al.,
  "Collaborative Heterogeneity-Aware OS Scheduler for Asymmetric Multicore Processors",
  TPDS'21 (https://ieeexplore.ieee.org/document/9296343)

5. Modern Batch Scheduling by example of FLUX

- Ahn, et.al.,
  "Flux: A Next-Generation Resource Management Framework for Large HPC Centers",
  ICPP 2014 (https://flux-framework.org/publications/Flux-SRMPDS-final.pdf)
- Ahn, et.al.,
  "Flux: Overcoming Scheduling Challenges for Exascale Workflows",
  WORKS 2018 (https://ieeexplore.ieee.org/abstract/document/8638377)
- Pollard, et.al.,
  "Evaluation of an Interference-free Node Allocation Policy on Fat-tree Clusters",
  SC'18 (https://flux-framework.org/publications/a26-pollard.pdf)

6. Clouds and Container Scheduling

- Singh and Chana,
  "A Survey on Resource Scheduling in Cloud Computing: Issues and Challenges",
  Journal of Grid Computing 2016, (https://link.springer.com/article/10.1007/s10723-015-9359-2)
- Meyer, et.al.,
  "Towards Interference-Aware Dynamic Scheduling in Virtualized Environments",
  JSSPP 2020 (https://link.springer.com/chapter/10.1007/978-3-030-63171-0_1)
- Ahmad, et.al.,
  "Container scheduling techniques: A Survey and assessment",
  J of King Saud University - Computer and Information Sciences
  (https://www.sciencedirect.com/science/article/pii/S1319157821000562?via%3Dihub)

7. Large-Scale Cluster management by example of Borg

- Verma, et.al.,
  "Large-scale cluster management at Google with Borg",
  EuroSys 2015 (https://research.google/pubs/pub43438/)
- Tirmazi, et.al.
  "Borg: the Next Generation",
  EuroSys 2020 (https://research.google/pubs/pub49065/)
- Rzadca, et.al.,
  "Autopilot: workload autoscaling at Google",
  EuroSys '20 (https://dl.acm.org/doi/10.1145/3342195.3387524)

8. Scheduling for Heterogeneous Systems

- Zimmer, et.al.,
  "GPU age-aware scheduling to improve the reliability of leadership jobs on Titan",
  SC'18 (`https://dl.acm.org/doi/10.1109/SC.2018.00010`)
- Ren and Tang,
  "Busy-Time Scheduling on Heterogeneous Machines",
  IPDPS'20 (`https://personal.ntu.edu.sg/asxytang/papers/ipdps2020.pdf`)
- Han, et.al.,
  "MARBLE: A Multi-GPU Aware Job Scheduler for Deep Learning on HPC Systems",
  CCGRID 2020 (`https://par.nsf.gov/biblio/10167822`)

9. Workflows and scheduling

- Rodrigo, et.al.,
  "Enabling Workflow-Aware Scheduling on HPC Systems",
  HPDC'17 (https://dl.acm.org/doi/10.1145/3078597.3078604)
- Terraz, et.al.,
  "Melissa: large scale in transit sensitivity analysis avoiding intermediate files",
  SC'17 (https://dl.acm.org/doi/10.1145/3126908.3126922)
- Hataishi, et.al.,
  "GLUME: A Strategy for Reducing Workflow Execution Times on Batch-Scheduled Platforms",
  JSSPP 2021, (https://link.springer.com/chapter/10.1007/978-3-030-88224-2_11)

10. Co-Scheduling

- Park, et.al.,
  "Hypart: a hybrid technique for practical memory bandwidth partitioning on commodity servers",
  PACT'18 (https://dl.acm.org/doi/10.1145/3243176.3243211)
- Aupy, et.al.,
  "Co-Scheduling HPC Workloads on Cache-Partitioned CMP Platforms",
  CLUSTER'18 (https://people.bordeaux.inria.fr/gaupy/ressources/pub/confs/cluster18_cache.pdf)
- Zhu, et.al.
  "Co-Run Scheduling with Power Cap on Integrated CPU-GPU Systems",
  IPDPS'17 (https://people.engr.ncsu.edu/xshen5/Publications/ipdps17.pdf)

11. IO-Aware Scheduling

- Fan, et.al.,
  "Scheduling Beyond CPUs for HPC",
  HPDC'19 (https://dl.acm.org/doi/10.1145/3307681.3325401)
- Herbein, et.al.,
  "Scalable I/O-Aware Job Scheduling for Burst Buffer Enabled HPC Clusters",
  HPDC'16 (https://dl.acm.org/doi/10.1145/2907294.2907316)
- Carretero, et.al.„
  "Mapping and scheduling HPC applications for optimizing I/O",
  ICS'20 (https://dl.acm.org/doi/10.1145/3392717.3392764)

12. Malleability

- Singh and Carretero,
  “Combining malleability and I/O control mechanisms to enhance the execution of multiple applications”
  J. Syst. Softw., 2019 (https://doi.org/10.1016/j.jss.2018.11.006)
- Comprés, et.al.,
  “Infrastructure and API Extensions for Elastic Execution of MPI Applications”,
  EuroMPI 2016 (https://doi.org/10.1145/2966884.2966917)
- D’Amico, Jokanovic and Corbalan,
  “Holistic Slowdown Driven Scheduling and Resource Management for Malleable Jobs”,
  ICPP 2019 (https://doi.org/10.1145/3337821.3337909)

13. Power-Aware Scheduling

- Acun, et.al.,
  "Power Aware Heterogeneous Node Assembly",
  HPCA'19 (https://ieeexplore.ieee.org/document/8675210)
- Chasapis, et.al.,
  "Power efficient job scheduling by predicting the impact of processor manufacturing variability,"
  ICS'19 (https://dl.acm.org/doi/10.1145/3330345.3330372)
- Wallace, et.al.,
  "A data driven scheduling approach for power management on HPC systems",
  SC'16 (https://dl.acm.org/doi/10.5555/3014904.3014979)

14. ML-Based Scheduling

- Zhang, et.al.,
  "RLScheduler: an automated HPC batch job scheduler using reinforcement learning",
  SC'20 (https://dl.acm.org/doi/abs/10.5555/3433701.3433742)
- Carastan-Santos and de Camargo,
  "Obtaining dynamic scheduling policies with simulation and machine learning",
  SC'17 (https://hal.inria.fr/hal-01618940/document)
- Fan, et.al.
  "Deep Reinforcement Agent for Scheduling in HPC",
  IPDPS'21 (https://arxiv.org/pdf/2102.06243.pdf)

15. Scheduling for Quantum Systems

- Humble, Britt and Mohiyaddin,
  "Quantum Accelerators for High-Performance Computing Systems",
  ICRC 2017 (https://www.osti.gov/biblio/1412048)
- Britt and Humble,
  "High-Performance Computing with Quantum Processing Units",
  JETC 2017 (https://doi.org/10.1145/3007651)
- McCaskey, et.al.,
  "XACC: a system-level software infrastructure for heterogeneous quantum–classical computing",
  Quantum Science and Technology 2020 (https://iopscience.iop.org/article/10.1088/2058-9565/ab6bf6)
- Khammassi, et.al.,
  "OpenQL: A Portable Quantum Programming Framework for Quantum Accelerators",
  JETC 2022 (https://dl.acm.org/doi/10.1145/3474222)

Topics

1. Metrics and Fundamentals of Scheduling
2. Threads and Tasks Scheduling
3. GPU Scheduling (Warps and Compute Units)
4. Process Scheduling (OS)
5. Modern Batch Scheduling by example of FLUX
6. Clouds and Container Scheduling
7. Large-Scale Cluster management by example of Borg
8. Scheduling for Heterogeneous Systems
9. Workflows and scheduling
10. Co-Scheduling
11. IO-Aware Scheduling
12. Malleability
13. Power-Aware Scheduling
14. ML-Based Scheduling
15. Scheduling for Quantum Systems

Next Steps:

- Tasks:
  - Mandatory Reading
    & research your topic
  - Get templates:
    Presentation and Paper
  - Make an outline for both your talk and paper!

- Next Session:
  Guidlines for:
  - Scientific writing
  - Presentations
- First Student Presentation:
  23.05.2022 (Topics 1. & 2.)

Backup

Links

Access to university library resources:

- `https://www.ub.tum.de/eaccess`

Template for IEEEtran:

- `https://www.ieee.org/conferences/publishing/templates.html`
- Use latex please!

Template for TUM presentation:

- `https://gitlab.lrz.de/latex4ei/tum-templates`
- `https://gitlab.lrz.de/matthias.maiterth/tum-templates`
- Any TUM Template! `http://go.tum.de/573475`
  You can use any presentation software you like.
  (Latex, Open Document Presentation, Keynote, PPT)