Cost-Efficient Job Scheduling Strategies

Alexander Leib

Arbeitsbereich Wissenschaftliches Rechnen
Fachbereich Informatik
Fakultät für Mathematik, Informatik und Naturwissenschaften
Universität Hamburg

2015-12-07

Agenda

Introduction

- Effectiveness
  - “Capability of producing or achieving a desired result” [5]
  - A means of achieving a set goal faster is more effective than other ones
  - The term is focused on the outcome
- Efficiency
  - "The ability to do things well, successfully, and **without waste**"[5]
  - Can be expressed as a measure of how well input is used to produce a desired output
  - Optimizing the **ratio** between input and output
  - E.g. in physics: Energy conversion efficiency
- Context HPC
  - Higher Performance and throughput can be more **effective** at the cost of **efficiency**.

Introduction

- Main expenses in HPC
  - Initial Investment
    - High costs for the aquisition of the machine(-components)
    - Suitable facilities must be availible or aquired
    - Electrical and network infrastructure
    - The initial investment is a one-time expense and relatively fixed
  - Running costs
    - Maintenance
    - Spare and replacement parts
    - Energy Consumption
    - Variable and regular expenses
- Cost reduction potential
  - The variability of the running costs results in a high saving potential by efficient running
  - Power consumption is probably the most flexible type of running costs

Introduction

Cost-efficiency = energy-efficiency?

- Cost-efficiency: Reduction of input (costs) at stable output (performance)
- Energy-efficiency: Reduction of Energy Consumption
- Power demand of HPC centers can reach 20 MW or more [1]
- High impact on electricity service providers and grid reliability
  - Power fluctuations
  - Peak power demands
- Intuitive assumption: Lower power consumption equals lower costs

Introduction

Electricity service provider (ESP)

- Goals
  - Prevention of transmission and distribution congestion
  - Frequency response
  - Peak and reserve capacity
  - Renewable integration
- Programs
  - Energy efficiency
  - Peak load reduction
  - Dynamic pricing
  - Up and down regulation

Introduction

HPC-Centers

- The HPC-landscape is changing
  - Instead of using higher performance components, systems use higher numbers of more energy-efficient components
  - Overall power demand is still rising
  - In compliance to ESP programs and/or requests, strategies to control electricity demand should be implemented
- Adaptation strategies
  - Power management
  - Load migration
  - Job scheduling
  - Back-up scheduling
  - Lighting control
  - Thermal management

Energy-efficiency

- Power consumption has been rising significantly slower than expected over the past years

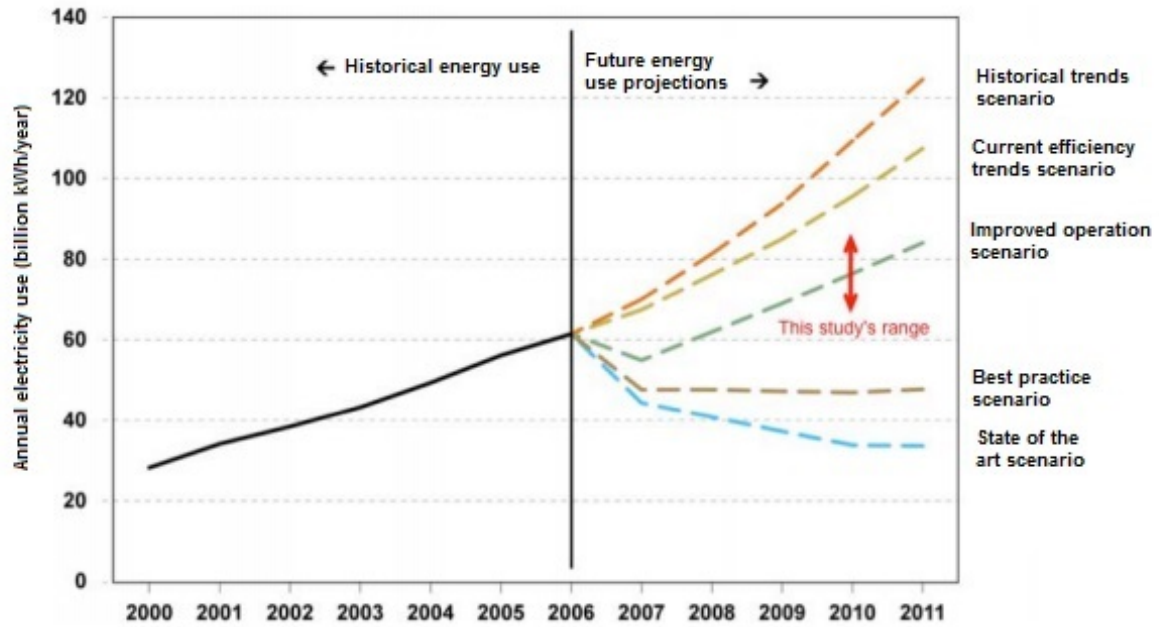

Abbildung: Energy useage in HPC

The question how this was achieved becomes apparent

Energy-efficiency

Achieving energy-efficiency in HPC-environments

- Energy-efficient or energy proportional hardware
  - The design of new hardware is not solely focused on higher performance, but also on optimizing the utilisation of power
  - By using energy-efficient components, the machines themselves become more energy-efficient
- Dynamic Power Management (DPM)
  - Devices (like CPUs, memory etc.) powered on and off dynamically
  - Idle machines consume about 2/3 of peak power and the average workload in data centers amounts to around 30%
  - => 70% of resources can be kept in sleep mode

Energy-efficiency

Power Management

- Dynamic Voltage and Frequency Scaling (DVFS)
  - Basic Idea: Adjustment of CPU clock frequency through supply voltage to reduce power consumption
  - Energy is consumed proportional to workload
- Power Capping
  - The operator can set a threshold for power consumption
  - Total power consumption can be kept under a defined budget
  - Sudden rises in power demand can be prevented
  - Achieved by i.e. throttling CPUs and de- or rescheduling tasks
- Thermal Management
  - Higher temperatures can reduce the system's reliability while increasing cooling expenses
  - Similar to Power Capping but the temperature is monitored and a threshold predefined

Energy-efficiency

- Highly parallel systems allow adjustment of power consumption to workload through job scheduling
- Job scheduling policies can be static or dynamic
  - Static: Jobs are known beforehand
  - Dynamic: Scheduling is done at job-arrival
- At the time of design energy efficiency has traditionally been a rare consideration, since the focus was on high throughput and performance

Energy-efficiency

- In most HPC-systems user requests are collected by a job scheduler
- The scheduler assigns jobs to nodes and specifies the time of execution
- The job scheduling algorithms offer a wide variety of opportunities to optimize the system for desired characteristics
- Optimizing the scheduler for energy-efficiency can drastically reduce power consumption
- In [2] Mämmelä et al. used different algorithms to compare results of different approaches

Energy-efficiency

Following algorithms were used in [2]

- First In, First Out (**FIFO**)
  - In this simplest of scheduling algorithms jobs are scheduled in a queue
  - If enough resources for the first job (job 1) are available it is executed, otherwise all jobs have to wait
- Energy aware FIFO (**E-FIFO**)
  - FIFO is expanded by a the power-off threshold T
  - If the estimated waiting time is longer than T seconds, idle nodes are powered off

Energy-efficiency

- Backfilling (first fit and best fit)
  - Backfilling basically works like FIFO with an important exception
  - If the available resources are insufficient for job 1, the queue is scanned for jobs that can be executed
  - The execution time of a potential backfill job must not exceed the waiting time of job 1 to prevent delays
  - Time estimations for job 1 are based on running time estimations of currently running jobs which are given by the user
  - User estimations can result in delays

Energy-efficiency

Two basic ways of selecting a backfill job are described in [2]

- Backfill first fit (**BFF**)
  - The first queued job meeting the resource and time restrictions is chosen
- Backfill best fit (**BBF**)
  - Additional criteria can be defined (i.e. shortest backfill, or most processors used)
  - Backfill job is selected accordingly

The energy-aware versions of backfilling (**E-BFF** and **E-BBF** use the same energy saving mechanisms as E-FIFO

- Idle nodes are powered off, if the wait time of job 1 exceeds T
- Backfilling offers less opportunities to power off idle nodes, by utilizing them for other jobs instead
- Wait times of jobs are however shorter than FIFO resulting in higher efficiency

Energy-efficiency

Two basic ways of selecting a backfill job are described in [2]

- Backfill first fit (**BFF**)
  - The first queued job meeting the resource and time restrictions is chosen
- Backfill best fit (**BBF**)
  - Additional criteria can be defined (i.e. shortest backfill, or most processors used)
  - Backfill job is selected accordingly

The energy-aware versions of backfilling (**E-BFF** and **E-BBF** use the same energy saving mechanisms as E-FIFO

- Idle nodes are powered off, if the wait time of job 1 exceeds T
- Backfilling offers less opportunities to power off idle nodes, by utilizing them for other jobs instead
- Wait times of jobs are however shorter than FIFO resulting in higher efficiency

Energy-efficiency

Simulation results using the 6 mentioned algorithms

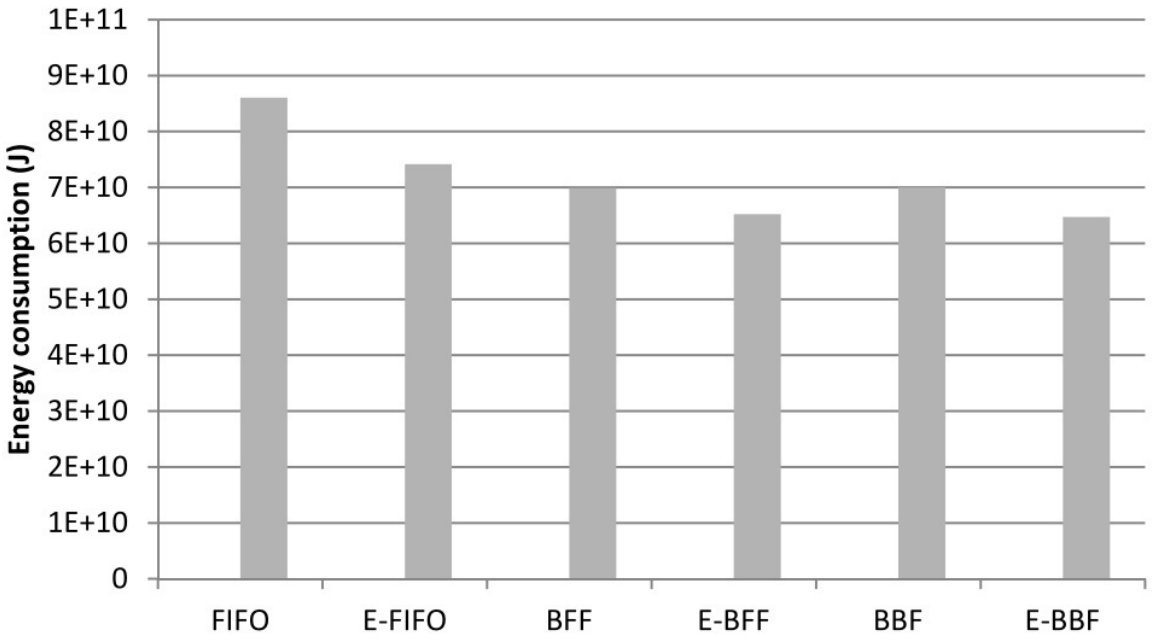

Abbildung: Energy Consumption

Wait and simulation times: BBF $<$ BFF $<<$ FIFO

Energy-efficiency

- Selected simulation results
  - The increase in average wait times of energy-aware scheduling algorithms did not exceed 3.2% (FIFO vs. E-FIFO tested with only small jobs requiring 1 - 5 nodes)
  - Highest average simulation time increase was 2.3% when comparing E-algorithms to their counterparts
  - Using a backfilling algorithm cut the simulation time by 23% compared to FIFO
- Results on a cluster at Jülich Supercomputing Centre (JSC)
  - JSC's testing environment uses the default scheduler of Torque RMS which is relatively close to BFF
  - Testing with the E-BFF algorithm resulted in a marginally higher (0.63%) completion time
  - In contrast: Energy consumption was 6.3% lower

Cost-efficiency

Approaches for working together with ESPs

- Flat Power Consumption
  - By "flattening" (i.e. keeping constant) the overall power consumption providers have a predictable power demand
  - Reduced fluctuations have less impact on electrical grid stability
  - Some ESPs offer flat fees for relatively constant power demand
- Dynamically Adjusted Power Consumption
  - Power demand regulation in response to grid capacity and ESP requests
  - ESPs often provide financial incentives to reduce power demand during peak times
  - By working closely together with providers, HPC-sites can actually be used as a resource to reduce the impact of grid fluctuations produced by other customers

Cost-efficiency

In [1] Bates et al. (2014) submitted a questionnaire to Supercomputing centers (SC) in the U.S. to evaluate the grade of interaction between SCs and ESPs

- 11 of 19 targeted U.S. Top100 SCs sent answers to the authors of [1]
- Results are more representative of the Top50, since 9 of 11 respondents are listed there
- SC's experiences concerning total power load and intra-hour fluctuations varied greatly, i.e.:
  - 4 sites with had a total power load of over 10 MW with fluctuations between $<$ 3 MW to 8 MW
  - Of two 5 MW sites, one reported 4 MW fluctuations

Cost-efficiency

Questionnaire results

- Approx. half the respondents have had some interaction/discussion, but mostly limited to ESP-programs like dynamic pricing
- About 20% have had discussions about congestion, regulation and frequency response

Discussions with ESPs

Discussions with ESPs	%Yes
Demand-side programs	
Shedding load during peak demand	54
Responding to pricing incentive programs	45
Shifting load during peak demand	36
Supply-side programs	
Enabling use of renewables	36
Congestion, regulation, frequency response	18
Contributing to electrical grid storage	10

HPC Strategies Responding to Electricity Provider Requests

HPC strategies for responding to electricity provider requests (listed from highest to lowest interest + impact)	% Interested	% High Impact	% Medium Impact
Coarse grained power management	64	46	27
Facility shutdown	36	64	10
Job scheduling	36	27	18
Load migration	10	36	18
Re-scheduling back-ups	45	0	10
Fine-grained power management	27	0	36
Temperature control beyond ASHRAE limits	27	0	18
Turn off lighting	18	0	0
Use back-up resources (e. g., generators)	0	10	27

Abbildung: Questionnaire results

Cost-efficiency

Interpretation of the results

- Due to local differences, these results are only representative of the U.S.
- The current low grade of interaction between SCs and ESPs has serveral reasons:
  - Organisational reasons (budgeting is done somewhere else)
  - Some power management strategies are considered to have negligible effects
  - The inherent trade-off in performance associated with most energy-efficience measures is deemed uneconomically high
  - Measures like thermal and lighting management do not seem to make much of a difference
  - The effects of highly sophisticated and automated job scheduling strategies is underestimated

Conclusion and evaluation

- Different means of reducing power consumption and combinations thereof can greatly enhance overall efficiency
  - Utilization of these approaches is surprisingly uncommon
  - Implementation of theses strategies may have to be realized on different levels
    - Top-down: Regulations by the decision makers of HPC-sites
    - Bottom-up: Initiatives by users
- Reaching understandings and closing deals with ESPs offer high cost saving potentials
  - The scope of possible actions is influenced by ESPs characteristics like grid infrastructure
  - Availible agreement opportunities may vary depending on the size and power demand of HPC-sites
    - Smaller sites may utilize flat power consumption strategies
    - SCs with higher demand my be forced to use dynamic strategies

Conclusion and evaluation

Cooperation with ESPs can result in lower prices for higher overall amounts of energy compared to purely energy-aware strategies

- Suggestions for compliance to ESP requests
  - Flat power consumption:
    - Fluctuations can be reduced at the job scheduler level
    - The wattages of different scheduler implementations differ
    - Power consumption can be influenced accordingly by switching algorithms
    - This process probably could be automated
  - Dynamic power consumption
    - Power consumption should be adjusted depending on current grid capacity and ESP requests
    - Can be achieved by utilizing more or less energy-efficiency means and scheduler algorithms

Literature

1. N. Bates et al.: Electrical Grid and Supercomputing Centers: An Investigative Analysis of Emerging Opportunities and Challenges
2. O. Mämmelä et al.: Energy-aware job scheduler for high-performance computing
3. A. Chandio et al.: A comparative study on resource allocation and energy efficient job scheduling strategies in large-scale parallel computing systems
4. D. Yang et al.: Parallel-machine scheduling with controllable processing times and rate-modifying activities to minimise total cost involving total completion time and job compressions
5. Wikipedia