



**Czech  
Technical  
University  
in Prague**

**F3**

**Faculty of Electrical Engineering  
Department of Computer Science**

## **Unassisted project report**

**Lukáš Forst**

**Supervisor: Ondřej Vaněk, Ph.D.  
January 2019**



# Contents

<b>1 Introduction</b>	<b>1</b>
<b>2 Problem definition</b>	<b>3</b>
2.1 Formal definition . . . . .	3
2.1.1 Variables Definition . . . . .	3
2.1.2 Resources reconfiguration . . . .	6
2.1.3 Optimization criteria . . . . .	6
<b>3 State of the art</b>	<b>7</b>
3.1 Load Balancing . . . . .	7
3.1.1 Static Load Balancing . . . . .	7
3.1.2 Dynamic Load Balancing . . . .	9
3.1.3 Load Balancing for Optimization Algorithms . . . . .	12
3.2 Optimization Algorithms . . . . .	13
3.2.1 Linear Optimization . . . . .	13
3.2.2 Heuristic algorithms . . . . .	14
3.2.3 Selected algorithms . . . . .	14
<b>Bibliography</b>	<b>17</b>



# Chapter 1

## Introduction

Optimization algorithms and solutions build on them are widely used in current manufacturing industry to reduce production costs. With more and more production automatization, optimization algorithms can manage and schedule whole factories with maximum available efficiency.

Complexity of optimization problems could be huge and therefore performance requirements are sometimes not easily satisfiable. Using one powerful instance of optimization algorithm in cloud seems like a solution for problems with smaller complexity, but what if we have multiple huge problems where each is performance demanding? Of course, we can create multiple instances, but that would be expensive and not well manageable and scalable since adding another instances manually requires some time and it is not much flexible. Another disadvantage of this approach is the fact, that optimization algorithm is not running 100% of time and thus resources allocated by this algorithm are unused while other algorithm instances could be potentially overwhelmed. Also paying for unused hardware is wasting money and optimization algorithms are supposed to save money.

Now imagine having two completely different problems that each requires its own application which visualises data and optimization algorithm to compute some kind of plan, this algorithm can be generic enough to operate on both domains with same code base, but it requires a lot of performance resources. If we use monolithic architecture of both applications, we would have same code in two applications, but what is even worse, we would need two powerful machines to run our applications. As previously mentioned, these two machines would not be using their power whole time and would be mainly idle.

What if one application runs only few minutes a day, but needs that power to complete tasks in time? A lot of resources would be wasted if it has its own server, but using not powerful server would lead to increasing duration of ongoing tasks which is something we do not want.

In this paper I would like to introduce **load balancer** specifically developed for optimization algorithms which could potentially minimize resources wasting and increase performance using correct utilization distribution across multiple instances of optimization algorithms.

Whole text does not seems to be right, maybe I will need to rewrite it.



## Chapter 2

### Problem definition

The problem with implementation of optimization algorithms in applications is that their performance requirements are quite high and are fully utilized only while working. Optimization algorithm is not running all the time and for that reason hardware resources are mainly unused. These unused resources could be potentially used by another instance of algorithm or can be shutdown completely to reduce hosting costs.

Also adding more time to job execution does not always bring better solution but it certainly costs more. Therefore proposed load balancer must be able to stop execution when solution value is not getting better in comparison with scheduling costs.

### 2.1 Formal definition

In the first place, used variables have to be introduced.

this must be changed

#### 2.1.1 Variables Definition

Used variables are derived into multiple categories: *indexes* - that are used to identify, or specify relation of variable, *input variables* - those are data provided to the system as an input from client application or user, *variables* - changed or computed during the runtime of the system.

Maybe change it a bit

Maybe add note about time representation in the system

#### Indexes

- $j$  - index used to identify something related to the execution job, in real world this is most likely job id, located in the right upper corner -  $x^j$ , set of all jobs in the system is represented by  $J$
- $r$  - index used to identify resources, written in the left upper corner -  $^rx$ , set of resources is represented by  $R$
- $t$  - right bottom index represents time -  $x_t$

### Input

Input which is specified before executing optimization job by the user outside of the system. When new execution job is requested - this is done by the user, or client application, following data should be provided.

- $D^j$  - maximal duration of the job execution which cannot be exceeded
- $P^j$  - maximal used resources cost per job, or in other words highest possible price paid for the job execution which cannot be exceeded

There are also constants defined before load balancer start up.

- $r_c$  - cost of the using particular resources per one time unit

Each of the previously mentioned variable must be non-negative.

### Program Output

Apart from result of the underlying optimization algorithm, following data are returned to user after successful job execution.

- $T^j$  - time taken, duration of the actual job execution
- $C^j$  - resource costs, how much job execution cost

### Variables

better example  
would be fine

- $v_t^j$  - value(i.e. cost of the scheduled plan) of the job  $j$  at the time  $t$  Value is greater than zero and it is non-increasing during the time.

$$\forall t, j : v_t^j \geq v_{t+1}^j > 0$$

It is non-increasing because optimization algorithms return always best found solution, so when worse solution, than currently best one, is found, returned is still the best solution found.

- $r x_t^j$  - represents assignment of the resources  $r$  at time  $t$  to job  $j$

$$r x_t^j = \{0, 1\}$$

Each  $x$  is either 1 = indexed resources are assigned to the job at given time or 0 = given combination does not have assignment. We assume that each job has only one such assignment at one time, which effectively means that this job is executed on the single computation node. This is defined by following constraint:

$$\forall j, t : \sum_{r \in R} r x_t^j = \{0, 1\}$$



- ${}^r\Delta_t^j$  - enhancement of the value  $v$  with resources  $r$  on the job  $j$  per time  $t$ .

$${}^r\Delta_t^j = {}^r|v_t^j - v_{t-1}^j|$$

It is improvement of the solution value  $v$  which can be achieved by using resources  $r$  at time  $t$ . This value is always non-negative since optimization algorithm always stores best found solution, and therefore  $\forall j, t, r : {}^r\Delta_t^j \geq 0$ .

- $S_t^j$  - reward for improving solution value in time  $t$  per job  $j$ . Accumulation of enhancements  ${}^r\Delta_t^j$  through all resources  $r$  and time units  $t$ .

$$S_t^j = \sum_{i=0}^t \sum_{r \in R} {}^r\Delta_i^j$$

- $C_t^j$  - defines how much execution of job  $j$  cost from the beginning of the execution until time  $t$ . Sum of all allocated resources for their time for the particular job.

$$C_t^j = \sum_{i=0}^t \sum_{r \in R} {}^rc \cdot {}^rx_t^j$$

Because  $C_t^j$  is defined as sum and  ${}^rc$  is non-negative, it is true that  $\forall j, t : C_{t+1}^j \geq C_t^j$ . Input of the program specifies maximal cost paid for job execution as a  $P^j$ , therefore it must be enforced by the system that this cost will not be exceeded. This constraint can be defined as follows.

$$\forall t, j : P^j \leq C_t^j \implies \sum_{i=t+1}^{\infty} \sum_{r \in R} {}^rx_i^j = 0$$

Which effectively means that when cost of job execution  $C_t^j$  has reached maximal defined cost  $P^j$ , no resources can be assigned to this job.

- $t$  - time, it is not only index but also variable, there are also constraints regarding time - since client application can specify deadline to job  $D^j$ , there must be additional constraint for job execution in a matter of resources assignment.

$$\forall t, j : D^j \leq t \implies \sum_{i=t+1}^{\infty} \sum_{r \in R} {}^rx_i^j = 0$$

When maximal time is exceeded, no additional resources can be assigned to the job execution, which could be defined by following constraint.

### 2.1.2 Resources reconfiguration

System should be capable of changing resources assignment per job in the runtime. This will help to distribute performance according to the current nodes load across whole network and allow to Unfortunately, it is not always possible to reconfigure resources assignment while scheduling is being performed. Therefore there must be at least one time unit, between different resources assignment, where no resources are assigned to the job. In other words, if resources reconfiguration is being done in time  $t$  then  $\sum_{r \in R} r x_t^j = 0$ .

This constraint can be defined in pseudocode, where  $m, n \in R$ ,  $m$  are assigned resources at the time  $t$ , this be written as  $m x_t^j = 1$ , and  $n$  are resources that should be assigned to the job in the time  $t + 1$ .

```

if  $m x_t^j = 1$  then
  if  $m = n$  then
     $n x_{t+1}^j = 1$ 
  else
     $\sum_{r \in R} r x_{t+1}^j = 0$ 
  end if
end if

```

Add mathematical definition of this constraint, is it even possible?

This constraint can defined as a mathematical function.

### 2.1.3 Optimization criteria

The main goal of the system is to minimize outcome value of the underlining optimization algorithm and at the same time to minimize cost of used resources. We can optimize single job or sum of outcomes from all jobs in the system at once. First approach provides possibility to control and optimize outcome of a particular job, which is an advantage for single client (owner o the job), but it does not necessarily means that it is optimal for the whole system and vice versa. Optimization criteria for the single job at particular time  $t$  is then maximization of weighted difference between the value enhancement reward and cost paid for the enhancement, which can be described by following equation.

$$\max crit_t^j = \alpha \cdot S_r^j - (1 - \alpha) \cdot C_t^j \quad 0 \leq \alpha \leq 1$$

For optimization of system-wide resources and costs, all jobs execution optimization is then defined like a weighted sum of all rewards per jobs lowered by sum of all resources costs across the set of all jobs.

$$\max crit_t = \alpha \sum_{j \in J} S_r^j - (1 - \alpha) \sum_{j \in J} C_t^j \quad 0 \leq \alpha \leq 1$$

## Chapter 3

### State of the art

#### 3.1 Load Balancing

Load balancing is technique for a division of processing work in the distributed environment of execution units <sup>1</sup> with aim to deliver faster service with higher efficiency. It improves the distribution of workloads across the whole environment and thus balances resources usage while maximizing throughput and minimizing response time. Load balancer is typically either dedicated *hardware device* or *software program*.

A **hardware** load balancer is a dedicated hardware device which distributes network traffic across a cluster of servers[Net]. These devices are used mainly in the data centers to ensure equal distribution of traffic between the application servers. Main benefit of using hardware load balancer is zero balancing overhead on the host machines, because all decisions are made on dedicated hardware specially developed for such tasks.

A **software** load balancer is a program operating on the application server with the same aim as hardware load balancer. Main advantage of the software load balancing is that it can be heavily customized and deployed to its own server. This paper will discuss only software load balancing approach.

In general, software load balancing algorithms can be classified as either *static* or *dynamic*.

##### 3.1.1 Static Load Balancing

Static load balancing is an approach where system information are provided a priori and load balancer does not use performance information about execution node <sup>2</sup>, to make distribution decisions. The performance possibilities and the load of the execution point (or node) are not taken in account when decision - where to execute current task - is being made, because load-balancing decisions are made at compile time. When a decision is made, no

---

<sup>1</sup>In general, execution unit can be CPU, network links, storage devices or other devices, in this paper *execution unit* or also referred as *execution node* or as *host* is a computer executing assigned job

<sup>2</sup>Execution node - Server executing task which is being scheduled by load balancer. In our case, this task is solving optimization problem by solver.

Find better example

other interaction with executing node, regarding the current task, is being made. In other words, once the load is allocated to the execution node, it cannot be transferred to another node. Static load balancing method is to reduce the overall execution time of a concurrent program while minimizing the communication delays[RP15]. The main advantage of static load balancing methods is mainly the fact, that there is minimal communication delay between system nodes and therefore execution overhead is minimized to almost zero. For that reason is static load balancing mainly used in the fields, where server response is crucial such as serving a web page. Also the implementation of some static load balancing algorithm is straightforward, since the used methods are very simple.

The main disadvantage of static load balancing is that it does not take in account current state of the system, when making decision. This could potentially lead to performance issues in the whole system because some nodes can be overloaded although others are not working at all.

Another drawback of this approach is that hardware resources are allocated only once in the execution time. Since optimization jobs are very heterogeneous, they sometimes have different power requirements during the execution. For example *TASP*<sup>3</sup> uses only one thread when creating feasible plan in the first algorithm iteration - this task relays only on single core performance. However, when first iteration is completed, all following can be done by multiple threads, therefore it could be useful to execute first iteration on a machine with better single core performance and then transfer algorithm into machine focused on multiple threads execution. This is something that can not be done while using static load balancing.

Following static load balancing algorithms are commonly used.

#### ■ First Alive

First alive or also called *Central Manager* algorithm uses the concept of a primary server and backup servers[IBM]. All tasks are scheduled to be executed on primary server unless the primary server is down. Then the load will be forwarded to first backup server. This algorithm has almost zero level of inner process communication, which leads to better performance when there are lots of smaller tasks.

#### ■ Round Robin

Round Robin algorithm which distributes work load evenly to all nodes. It is being done in round robin order, where load is distributed to each node in circular order without any priority. Round Robin is easy to implement and as well as *First alive* algorithm has almost none inner communication overhead. This algorithm performs best when tasks have equal, or at least similar, processing time.

<sup>3</sup>**Task and Asset Scheduling Platform** - proprietary optimization software developed by Blindspot Solutions, described in 3.2.2

### ■ Weighted Round Robin

Weighted round robin algorithm maintains a weighted list of servers and forwards new connections in proportion to the weight, or preference, of each server. This algorithm uses more computation times than the round robin algorithm. However, the additional computation results in distributing the traffic more efficiently to the server that is most capable of handling the request[IBM].

### ■ Threshold Algorithm

Threshold algorithm - execution nodes keep private copy of the system's load, when the load state of a node exceeds a load level limit, node sends message to all remote nodes, that it is overloaded. If the local state is not overloaded then the load is allocated locally. Otherwise a remote node, that is not overloaded, is selected and if no such node exists it is also allocated locally. This algorithm has low inter process communication and large number of local process allocations. The later reduces the overhead of remote process allocation and the overhead of remote memory access, which leads to performance improvements[PB].

### ■ Least Connections

Least connections algorithm maintains a record of active server connections and forward a new connection to the server with the least number of active connections[IBM]. This can be generally useful while having many concurrent requests, that can be dispatched quickly.

### ■ Randomized Algorithm

Randomized algorithm uses random selection of the execution node without having any information about it.

## ■ 3.1.2 Dynamic Load Balancing

Unlike static load balancing algorithms, dynamic algorithms use runtime state information to more informative decisions while distributing the jobs. They monitor changes on the system work load and take it in account when decision, where to execute job, is being made. The process of monitoring the system is not stopped after execution job started and if circumstances change, job execution can be transferred to another system node which then proceeds with execution.

While many different load balancing algorithms have been proposed, there are four basic steps that nearly all algorithms have in common[Mal00].

1. Monitoring workstation performance (load monitoring)
2. Exchanging this information between workstations (synchronization)

3. Calculating new distributions and making the work movement decision (rebalancing criteria)
4. Actual data movement (job migration)

Dynamic load balancing algorithms can be divided into two groups based on their control form, or in other words, where load balancing decisions are made [Mal00].

- Centralized - a single node in the network is responsible for all load distribution
- Distributed - all nodes are equal

While in centralized scheme decisions are made in one master workstation, in distributed scheme, the load balancing algorithm runs on all nodes and each node balances itself. Each of these approaches has its own ups and downs, centralized scheme can be a potential performance bottleneck since it relies on one system node, on the other hand distributed scheme has communication overhead, because it requires broadcast communication between all algorithm instances.

The main advantage of dynamic load balancing is that it allows changing execution node in runtime. For that reason it is possible to change hardware characteristics according to the job execution phase. For example execute initial phase of optimization algorithm on machine with powerful single core performance and then move the job to the machine with multiple, less powerful, cores to let it run in parallel. Also as a result of runtime scheduling, dynamic load balancing algorithms tend to provide significant improvements in performance over static algorithms. However, this comes at the additional cost of collecting and maintaining load information [Mal00]. For that reason dynamic load balancing suits better for long running tasks, which can be managed and distributed better, than for fast queries.

This sentence sounds weird

### ■ Dynamic load balancing strategies

There are three major parameters which usually define the strategy a specific load balancing algorithm will employ. These three parameters answer three important questions [Mal00]:

1. Who makes the load balancing decision?
2. What information is used to make the load balancing decision?
3. Where the load balancing decision is made?

Question number 1 is answered based on whether a **sender-initiated** or **receiver-initiated** policy is employed. In *sender-initiated* policies, congested nodes attempt to move work to lightly-loaded nodes. In *receiver-initiated* policies, lightly-loaded nodes look for heavily-loaded nodes from which work may be received [Mal00].

Question ‘What information is used to make the load balancing decision?’ is answered by following policies - **global** and **local**. When algorithm uses *global* policy, the load balancer uses the performance profiles of all execution nodes connected to the network. When using *local* policy, only local<sup>4</sup> nodes are taken in account while creating performance profile of the system.

The last parameter - ‘where the load balancing decision is made’ - is answered by used control form, as mentioned previously, dynamic load balancing algorithms are divided into two groups based on their control form - **centralized** and **distributed**.

I would like to present two general dynamic load balancing algorithms - *Central Queue Algorithm* and *Local Queue Algorithm*.

### ■ Central Queue Algorithm

Central queue algorithm is based on centralized receiver-initiated load balancing strategy. It uses a cyclic FIFO queue on the main host to store new activities<sup>5</sup> and unfulfilled requests. New activity request is inserted into queue and here it is stored until some execution node picks it up.

Whenever a request for an activity (which is send by executing node in the case when its load has fallen bellow specified threshold) is received by the queue manager<sup>6</sup>, it removes the first activity from the queue and sends it to the requester. If the queue is empty, the request is buffered, until a new activity is available. If a new activity arrives at the queue manager while there are unanswered requests in the queue, the first such request is removed from the queue and the new activity is assigned to it.

When a execution node load falls under the threshold, the local load manager sends a request for a new activity to the central load manager (which manages the central system queue). The central load manager answers the request immediately if a ready activity is found in the queue, or queues the request until a new activity arrives[SSS08].

### ■ Local Queue Algorithm

Local queue algorithms uses distributed receiver-initiated strategy.

Its main feature is, that it supports dynamic process migration. This algorithm in the first step uses static allocation of all new processes - all processes are allocated to under loaded hosts. In the second step the process migration is initiated by a host when its load falls under predefined threshold<sup>7</sup>. In such case, the execution node attempts to get several processes from remote hosts. It randomly sends requests with the number of local ready processes to remote load managers. When a load manager receives such a request, it

Weird sentence

<sup>4</sup>Workstations are usually divided into groups, in this context *local* means in the same group of workstations

<sup>5</sup>Activities - jobs to be executed, in our case optimization job

<sup>6</sup>Queue manager - central server which manages queue

<sup>7</sup> This threshold can be defined by the user and it is an input for the algorithm





## 3.2 Optimization Algorithms

This work does not contain any own algorithm implementation for generic optimization problems, instead I would like to use pre-prepared and already implemented optimization solver. We have many options how to solve optimization problems, I would like to present two of them - linear optimization and heuristics algorithms.

This is really shitty introduction

### 3.2.1 Linear Optimization

Linear optimization (or linear programming) is a method to achieve the best outcome in a mathematical model whose requirements are represented by linear relationships [Wik19]. The algorithms are widely utilized in company management, such as planning, production, transportation, technology and other issues.

I must add more about it since it is important topic

The main benefit of linear optimization is that it provides the best possible solution, because optimization algorithms are guaranteed to provide optimal solution. Although almost everything can be represented as linear problem, linear programming solvers could be unable to provide solution since, in the most cases, computation time grows exponentially. Even though there are solvers that are able to provide  $\epsilon$  (partial) solution, this solution can be (and in most cases is) unusable, because it is not optimal at all.

There are plenty of linear programming solvers available. I would like to highlight following two optimization kits.

#### GLPK

**GNU** - *GNU Linear Programming Kit* is a software package intended for solving large-scale linear programming (LP), mixed integer programming (MIP), and other related problems. It is a set of routines written in ANSI C and organized in the form of a callable library [Mak]. Although originally is GLPK written in C programming language, there is an independent project, which provides Java-based interface for execution of GLPK via Java Native Interface.<sup>8</sup>

#### Google OR-Tools

**Google OR-Tools** - OR-Tools is an open source software suite for optimization, tuned for tackling the world's toughest problems in vehicle routing, flows, integer and linear programming, and constraint programming [Goo]. Tools contain *Glop* which is Google's custom linear solver. One of the greatest advantages of Google OR-Tools is great API supporting multiple programming languages - C++, Python, C# and Java.

<sup>8</sup>Java Native Interface - Interface provided by Java platform to run and integrate non-Java language libraries



While choosing suitable solvers I was looking mainly at possibility running on JVM and their API as well as at their suitability for my paper. For final testing I selected **GLPK** as linear solver, mainly because it is widely used linear optimization kit and because of it's convenient Java interface.

As a representative of heuristics algorithms I selected **TASP** because of it's great scalability, Kotlin interface and because I have already worked with it and I'm familiar with multiple TASP implementations.

do I have to mention that I'm working for Blindspot?





## Bibliography

- [Goo] Google, *About or-tools*, <https://developers.google.com/optimization/>, [Online; accessed 16-January-2019].
- [IBM] IBM, *Algorithms for making load-balancing decisions*, [https://www.ibm.com/support/knowledgecenter/SS9H2Y\\_7.0/com.ibm.dp.doc/lbg\\_algorithms.html](https://www.ibm.com/support/knowledgecenter/SS9H2Y_7.0/com.ibm.dp.doc/lbg_algorithms.html), [Online; accessed 29-January-2019].
- [Mak] Andrew Makhorin, *Glpk (gnu linear programming kit)*, <https://www.gnu.org/software/glpk/>, [Online; accessed 16-January-2019].
- [Mal00] Malik, Shahzad, *Dynamic load balancing in a network of workstations*, 95.515 F Research Report (2000).
- [Net] AVI Networks, *Hardware load balancer*, <https://avinetworks.com/glossary/hardware-load-balancer/>, [Online; accessed 30-January-2019].
- [Pap18] Papanikolaou, A., *A Holistic Approach to Ship Design: Volume 1: Optimisation of Ship Design and Operation for Life Cycle*, Springer International Publishing, 2018, 296-301.
- [PB] Atul Garg Payal Beniwal, *A comparative study of static and dynamic load balancing algorithms*, International Journal of Advance Research in Computer Science and Management Studies, [Online; accessed 29-January-2019].
- [RP15] Dr. Samrat Khanna Ramesh Prajapati, Dushyantsinh Rathod, *Comparison of static and dynamic load balancing in grid computing*, International Journal For Technological Research In Engineering (2015).
- [SSS08] Sandeep Sharma, Sarabjit Singh, and Meenakshi Sharma, *Performance analysis of load balancing algorithms*, World Academy of Science, Engineering and Technology **38** (2008), no. 3, 269–272.
- [Wik19] Wikipedia contributors, *Linear programming — Wikipedia, the free encyclopedia*, <https://en.wikipedia.org/w/index.php?title=>

Linear\_programming&oldid=878407127, 2019, [Online; accessed 16-January-2019].