TuneUp.ai LightningCloud Performance Tuning Products Technical White Paper

April 16, 2019

What is Parameter Tuning

LightningCloud tunes the parameters of a system to optimize its performance. What is parameter tuning?

> “Choosing parameter values that suit your software, hardware, and workloads could greatly improve your systems’ performance.”

We all love fast systems

Of course, faster systems are easily translated to higher user satisfaction, less hardware, or lower cost of ownership. And **parameter tuning is one of the most effective ways to improve a system’s performance.**

All computer and electronic systems have tunable parameters, so that they could be adapted to different scenarios and workloads easily without changing the source code. Tuning these parameters according to your specific use cases and workloads could greatly increase your system’s performance. For instance, a video streaming website could face radically different workloads, such as when millions of users are watching the same breaking news or when each of them is watching a different movie. These different use cases require the system to have different settings for its I/O buffer size, worker thread count, network packet size, among hundreds of parameters.

Samples of Parameters

Microsoft recommends tuning the receive and send buffer sizes of the network adapters for intensive workloads on Windows systems. On Linux systems, the I/O scheduler and virtual memory settings should be chosen according to application’s I/O pattern.

However, finding the optimal values for these parameters require expertise and can be time consuming.

Default parameter settings is for safety, not performance

It is common knowledge that the default settings for most parameters are pretty conservative, so that the system could be set up easily and support most workloads out-of-

tuneup.ai
intelligent tuning

the-box. However, it is common for a computer system to receive a performance boost from 30% to several times after being tuned by the vendor experts.

Finding optimal parameter values is hard

Modern computer systems are complex behemoths with billions of transistors, and the software we are running can easily contain millions of line of code. Not to mention that each device usually needs to interact with other devices through a network, making their workloads even more complex.

The optimal parameter values depend on a wide range of factors, including but not limited to:

- Hardware configuration
- BIOS settings
- Software (OS, middleware, applications, etc., and their versions and patch levels)
- Network connection (speed, quality, topology, etc.)
- Workloads

Apparently, every computer system is different. Therefore, they would need to be tuned separately.

Finding optimal values for the parameters of a system requires a deep understanding of the internals of the target systems, including all aspects of hardware, network, software, and application, as well as the properties of the workloads running on the system. Therefore, parameter tuning is a service that usually only offered by the product vendor or experienced experts, both of which could easily be more expensive than the cost of the hardware and software themselves. And moreover, even with the best experts on site, they still need time to measure and understand the user's specific workloads, and this process could take days to weeks.

Parameter Tuning vs. Other Tuning Methods

Parameter tuning only changes the values of system parameters or configuration settings. It does not change hardware, alter network topology, adding/remove software modules, updating software versions, or modify the software source code. Parameter tuning usually only modify those safe parameters that would not affect the safety of user data.

The following table compares parameter tuning with other performance optimization methods.

	Changing system parameters	Upgrading hardware/software	Altering network topology	Modifying source code
When can be done	Any time	Needs planned downtime	Needs planned downtime	Needs planned downtime
Cost	Low to none	High	High	Very high
Affecting stability	Low	High	High	Very high
Downtime	None	Long	Medium	Long
Handling abrupt change to workload	Yes	Cannot	Cannot	Cannot
Risk	Low	High	Medium	Very high

Changing system parameters is usually easy and low-cost. However, the most challenging task is how to find out the optimal parameter values.

Who Needs Parameter Tuning

1. High traffic websites
2. Interactive systems
3. High performance database
4. Network appliances
5. Supercomputer
6. Data centers
7. Enterprise server and storage clusters

> "We didn't know our systems were under-performing until we saw a 40% boost in performance after tuning its parameters."

Challenges of Manual Parameter Tuning

Manual parameter tuning faces at least the following challenges:

1. Requires a deep understanding of the internals of the hardware, software, network, and applications, which usually requires coordinated efforts of vendor experts and the customer.
2. Usually requires a long downtime, during which multiple benchmarks have to be run to understand the workloads and measure the effect of tuning.
3. Is usually only done when the system was initially deployed and could not handle rapidly or slowly changing workloads.
4. Effect is highly unpredictable. User could only know about the result of tuning when it is actually done. There is no way to know the effect before spending money to hire the experts and doing the tuning.
5. Usually needs to be redone when there is a change to hardware, software, or the application.
6. Lack of transparent feedback. It is hard to measure if changing certain parameters actually has a positive effect on the desired performance or if the change in performance is stable or merely a fluke, especially when the change to performance is small or the workload itself is changing or periodical. To solve this issue, constant and costly monitoring and analysis are usually required.

> “Manual parameter tuning usually requires days or weeks to carry out.”

TuneUp.ai LightningCloud

TuneUp.ai LightningCloud is an easy-to-use fully automatic parameter tuning solution based on the latest machine learning and artificial intelligence research. LightningCloud uses **Deep Reinforcement Learning (DRL)**. We adapted the latest development of DRL for performance tuning, adding **patented methods to address its shortcomings** such as the demand for huge amount of training data, model instability, and overfitting

Deep Reinforcement Learning

Deep Reinforcement Learning is one of the latest development in deep learning research. Deep Reinforcement Learning enables a smart agent to explore and measure an unknown environment in a fully autonomous manner without any human supervision, and it has been shown to be able to outperform human in complex tasks such as Chess, Go, and Starcraft game playing, as well as data center cooling optimization.

over long term training. We also **fine-tuned our hyper-parameters** using our huge data sets accumulated over the years.

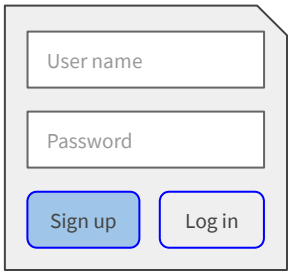
We published a paper detailing an early version of our system in Supercomputing '17. We were able to achieve a **45% increase** in throughput for Lustre without any manual tuning or supervision. Our latest DRL solution outperforms the prototype used in our 2017 paper by a wide margin.

Three Steps to Bring the Latest AI Tuning to Your System

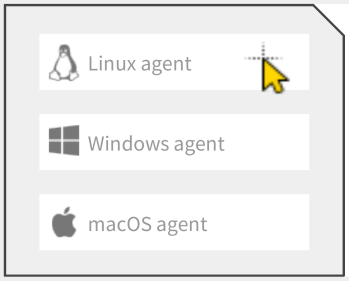
We see one of the biggest hurdles for the adoption of AI in the real world is the complexity of its deployment (anyone who has tried to set up a machine learning environment with GPU support could testify). We solved this problem by offering to run our artificial intelligence code in the cloud.

A customer only needs the following three steps to bring the latest AI performance tuning technologies to their systems:

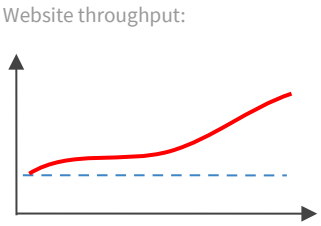
1. Register an account on https://tuneup.ai



2. Download and install a light-weight agent



3. Watch performance increases



Our Business Model

Our simple and elegant solution is made possible by the following three core features:

1. Machine learning in the cloud

All machine learning tasks are done in the cloud.

2. Cloud price model

Flexible Price Plans

- Free trial period.
- No need to pay if you are not satisfied with the result.
- Pay by how much you use.
- Different rates based on how many computers and parameters you tune.

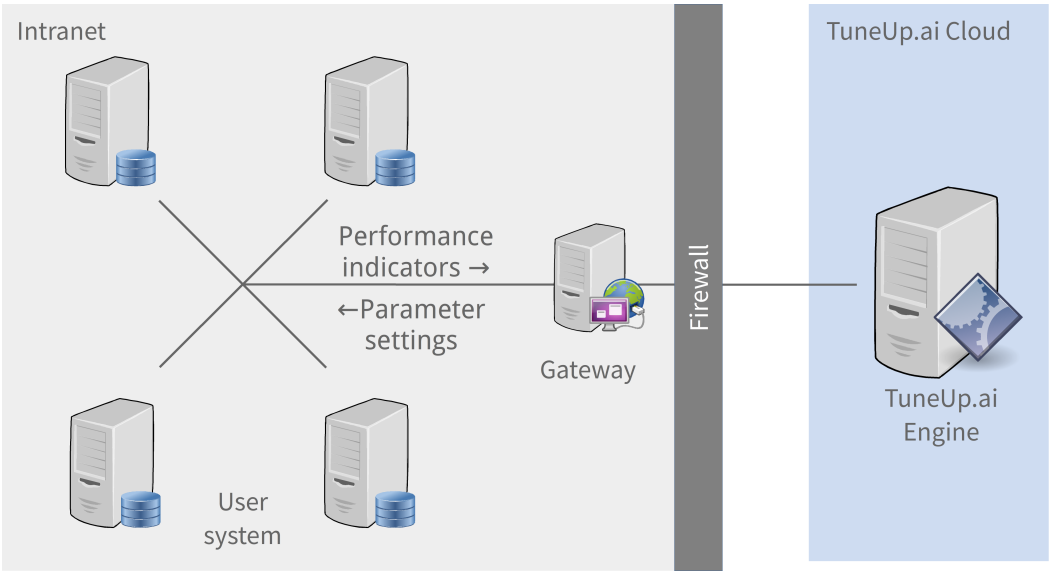
You pay for how much you use, and only when you are satisfied with the tuning result.

3. Fully open source agent

You control what you upload and what parameters to change with easy to expand Python scripts.

Machine Learning in the Cloud

Following diagram shows the deployment of LightningCloud:



The user only needs to deploy agents onto their systems and configure the tuning goal. Spinning up the engine and doing the machine learning are all automatically handled by our cloud infrastructure.

tuneup.ai
intelligent tuning

Automatic Performance Tuning vs. Manual Tuning

In addition to being able to monitor and tune the target system 24x7, our AI tuning system has many advantages over traditional manual parameter tuning methods:

	TuneUp.ai AI Tuning	Manual Parameter Tuning
Skill requirement	Regular IT staff	Vendor and domain experts
Total cost	Low and predictable	High and unpredictable
Applicable time	24x7x365, no downtime needed	Require planned downtime
Required time	10 hours to a few days	Days or weeks
Tuning effect	Most workloads will see improved performance	Depends on the skills of the experts
Handling rapidly changing workloads	Real time	Depends on response speed of staff
Safety	High	Depends on the process and staff skill
Handling hardware or software change	Automatically	Need to redo the tuning

Comparing with Other Machine Learning-based Methods

We have focused our research on using machine learning for performance tuning for many years and have developed several unique methods that have the following advantages when compared to other machine learning-based performance tuning methods:

	TuneUp.ai AI Tuning	Other machine learning implementations
Training process	Does not require a separate training process	Usually require a dedicated training process
Overfitting	Automatically measured and handled by our patented training engine	Need manual monitoring and mitigation
Setup process	Automatically handled in the cloud	Require multiple components and a complex setup process
Scalability	Can scale out to support as many machines as you need	Need capacity planning, and downtime is usually needed for scaling out
Coding effort	No need for common software. Customer can write scripts for collecting data and setting parameters for their proprietary systems.	A length development process is needed to develop software to collect data and setting parameters even if you use commercial Reinforcement Learning cloud platforms
Amount of training data	We have fined tuned our training engine to require as little data as possible before reaching optimal performance	Deep Reinforcement Learning could require millions of data points before it could converge

Flexible Price Model

Having the machine learning part running in the cloud also enables us to adopt a flexible price model. The following is a sample:

Basic Plan

- $0.2/h with 50 hours free trial period.
- No need to pay if you are not satisfied with the result.
- Credit card or deposit.
- Up to 5 hosts.
- Forum and community support.

Enterprise Plan

- $0.15/h with 100 hours free trial.
- No need to pay if you are not satisfied with the result.
- Credit card or deposit.
- 5 to 100 hosts.
- Email support. Response in two business days.

Data Center Plan

- $0.1/h with 200 hours free trial.
- No need to pay if you are not satisfied with the result.
- Credit card or deposit.
- 100 or more hosts.
- Phone/email support. Response in 24 hours.

Plans above are only samples. Actual terms are subject to change.

The Open Source, Extensible Client Agent

The LightningCloud Client Agent collects information from the target system, sends them to the machine learning engine in the cloud for analysis, and sets parameters according to the instructions from the machine learning engine. The agent is fully open source (under LGPL v2.1) so that users could examine its

source code and understand what information about their systems is sent to the cloud. It is written in Python and is extensible through extensions. Users could develop their own extensions to collect information from any systems and set any parameters.

Supported systems

The LightningCloud agent currently supports collecting performance information and tuning parameters of the following systems:

- Linux system
- NGINX
- Lustre

Collecting data from and setting parameters for other systems can be easily implemented in Python for our open source agent. We are also actively engaging with our customers to add more extensions to our agent.

About TuneUp.ai

TuneUp.ai is located in the San Francisco Bay Area. We focus on delivering the latest development of Artificial Intelligence technologies as easy-to-use solutions and applying them to solve systems problems. We maintain long term collaborations with top research labs in the world.

TuneUp.ai

https://tuneup.ai

Version: 4/16/2019

tuneup.ai
intelligent tuning