The 4th Innovating the Network for Data Intensive Science (INDIS) workshop

Towards a Smart Data Transfer Node

Zhengchun Liu, Rajkumar Kettimuthu, Ian Foster, Peter H. Beckman

Presented by: Zhengchun Liu

November 12, 2017, Denver CO

Argonne NATIONAL LABORATORY

Motivation

Motivation

Computer systems are getting ever ***more sophisticated***, and ***human-lead*** empirical-based approach towards system optimization is *not the most efficient* way to realize the full potential of these modern and complex high performance computing systems.

Motivation

Computer systems are getting ever ***more sophisticated***, and ***human-lead*** empirical-based approach towards system optimization is *not the most efficient* way to realize the full potential of these modern and complex high performance computing systems.

- ◉The effectiveness of parameters are not straightforward or intuitive understandable.
- ◉The system is dynamic. Fairly impossible to design a one-size-fits-all rule.
- ◉Parameter space is very big and very time consuming to explore.
- ◉Environment and platform are different.

Motivation

Computer systems are getting ever ***more sophisticated***, and ***human-lead*** empirical-based approach towards system optimization is *not the most efficient* way to realize the full potential of these modern and complex high performance computing systems.

- The effectiveness of parameters are not straightforward or intuitive understandable.
- The system is dynamic. Fairly impossible to design a one-size-fits-all rule.
- Parameter space is very big and very time consuming to explore.
- Environment and platform are different.

The data transfer nodes (DTN) are compute systems dedicated for wide area data transfers in distributed science environments.

Motivation

Computer systems are getting ever ***more sophisticated***, and ***human-lead*** empirical-based approach towards system optimization is *not the most efficient* way to realize the full potential of these modern and complex high performance computing systems.

◉The effectiveness of parameters are not straightforward or intuitive understandable.

◉The system is dynamic. Fairly impossible to design a one-size-fits-all rule.

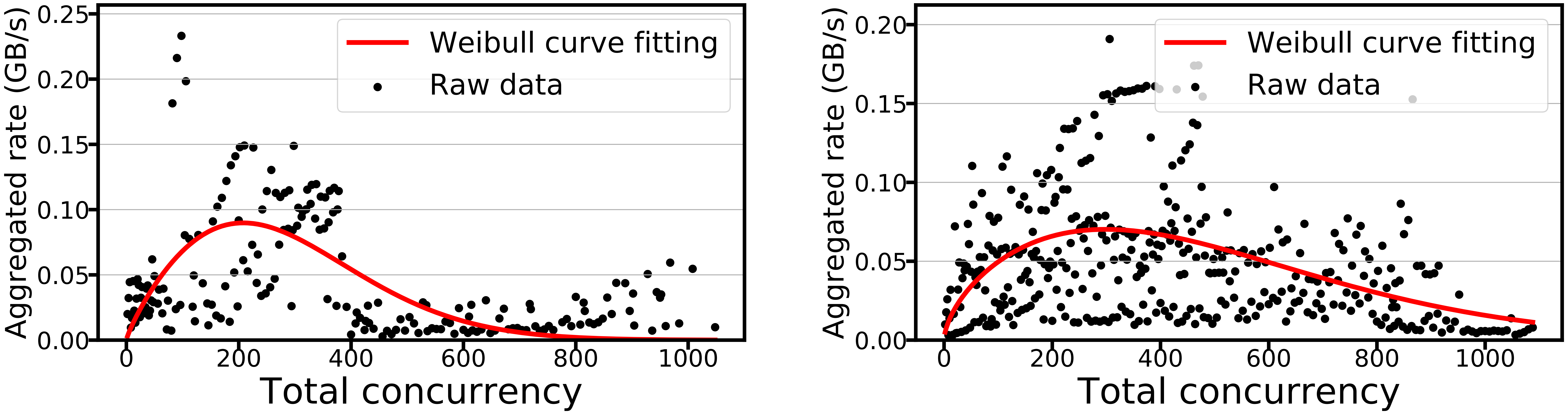
◉Parameter space is very big and very time consuming to explore.

◉Environment and platform are different.

The data transfer nodes (DTN) are compute systems dedicated for wide area data transfers in distributed science environments.

Inspired by work from Google Deepmind about using reinforcement learning to play games (e.g., AlphaGo, Atari). We use reinforcement machine learning methods to discover the ***"just right***" control parameters for data transfer nodes in dynamic environment.

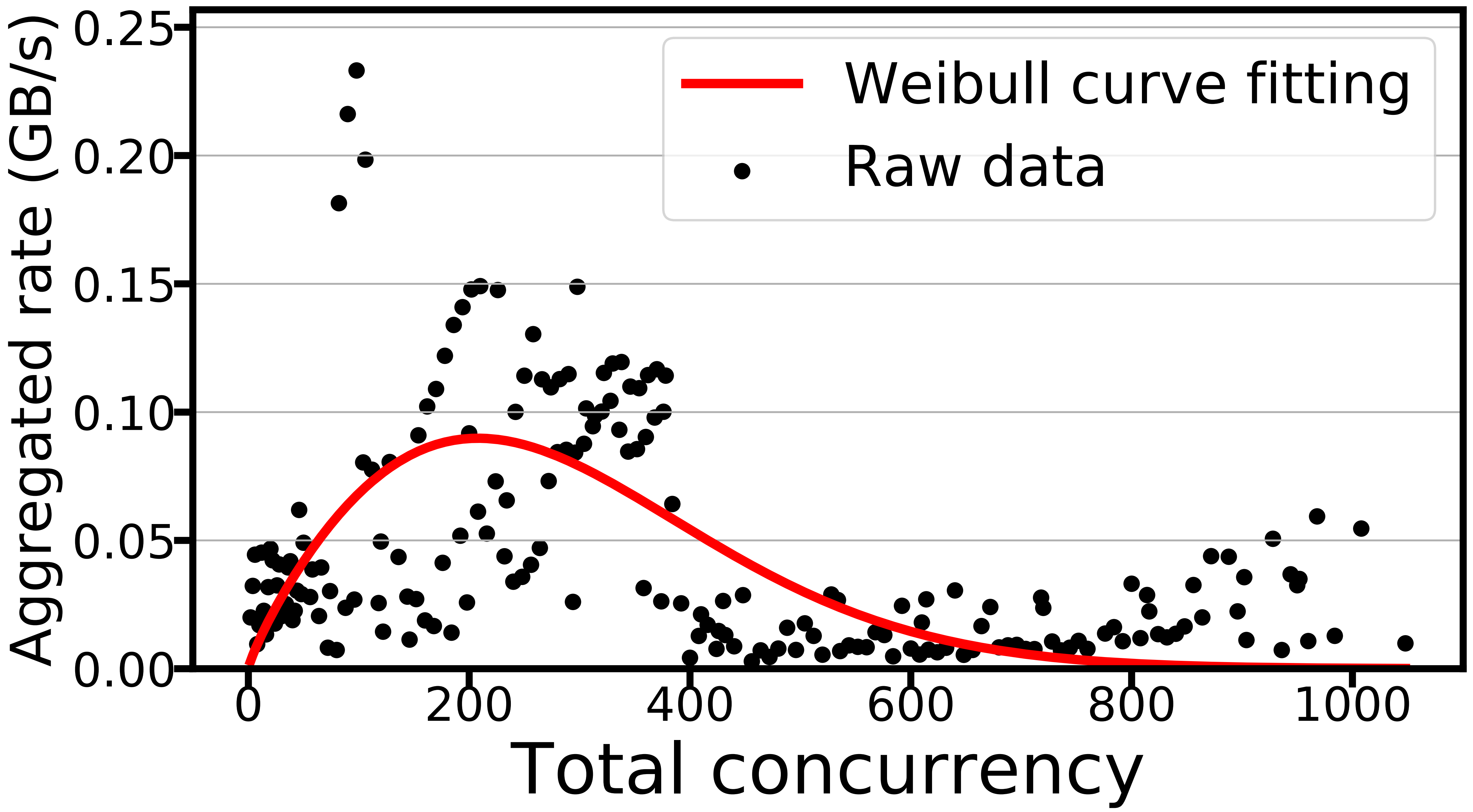
Motivation

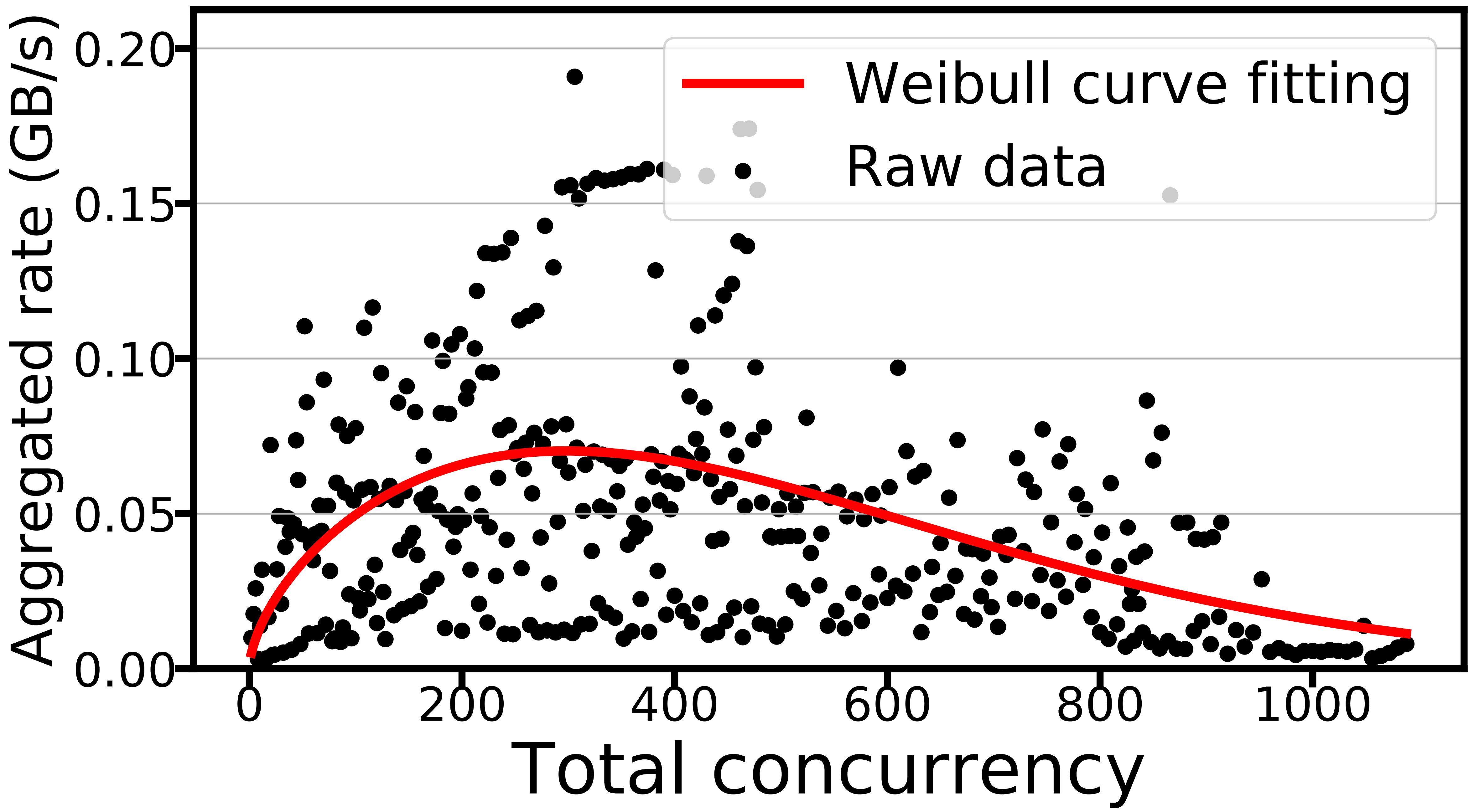


*** Aggregate incoming transfer rate vs. total concurrency (i.e., instantaneous number of GridFTP server instances) at two heavily used endpoints, with Weibull curve fitted.**

* Z. Liu, P. Balaprakash, R. Kettimuthu, I. Foster, Explaining wide area data transfer performance. HPDC'17

Motivation





*** Aggregate incoming transfer rate vs. total concurrency (i.e., instantaneous number of GridFTP server instances) at two heavily used endpoints, with Weibull curve fitted.**

Luckily, the optimal operating point of these two endpoints are almost fixed. However, the optimal operating point of most endpoints are dynamical because of continuously changing external load.

* Z. Liu, P. Balaprakash, R. Kettimuthu, I. Foster, Explaining wide area data transfer performance. HPDC'17

Reinforcement Learning

[Idea] An agent interacting with an environment, which provides its ***current state*** and numeric ***reward*** signals after each action the agent takes.

[Goal] ***Learn*** how to ***take actions*** in order to ***maximize reward.***

https://en.wikipedia.org/wiki/Q-learning

Reinforcement Learning

[Idea] An agent interacting with an environment, which provides its ***current state*** and numeric ***reward*** signals after each action the agent takes.

[Goal] ***Learn*** how to ***take actions*** in order to ***maximize reward***.

Agent (Controller)

https://en.wikipedia.org/wiki/Q-learning

Reinforcement Learning

[Idea] An agent interacting with an environment, which provides its ***current state*** and numeric ***reward*** signals after each action the agent takes.

[Goal] ***Learn*** how to ***take actions*** in order to ***maximize reward***.

Agent (Controller)

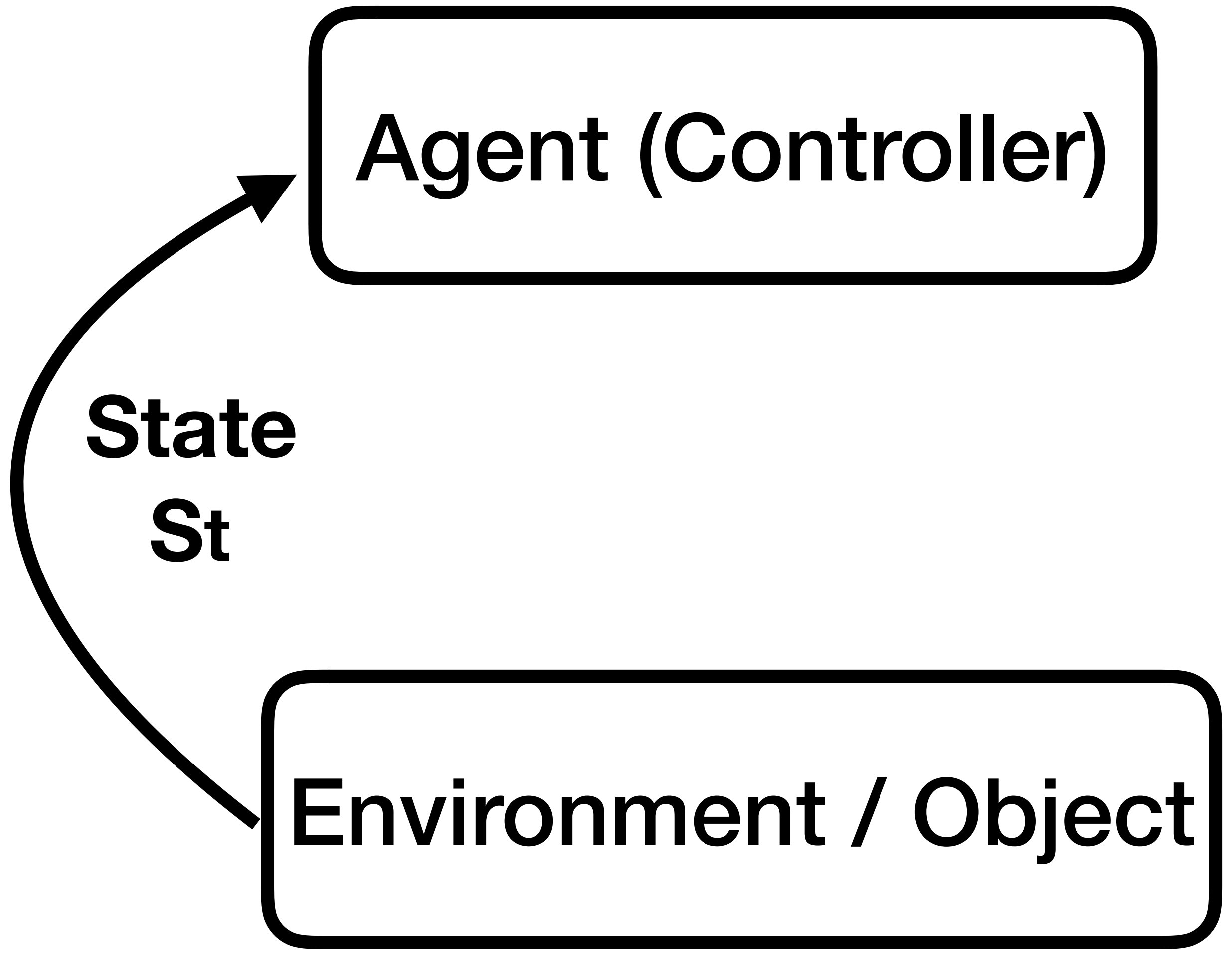
Environment / Object

https://en.wikipedia.org/wiki/Q-learning

Reinforcement Learning

[Idea] An agent interacting with an environment, which provides its ***current state*** and numeric ***reward*** signals after each action the agent takes.

[Goal] ***Learn*** how to ***take actions*** in order to ***maximize reward***.

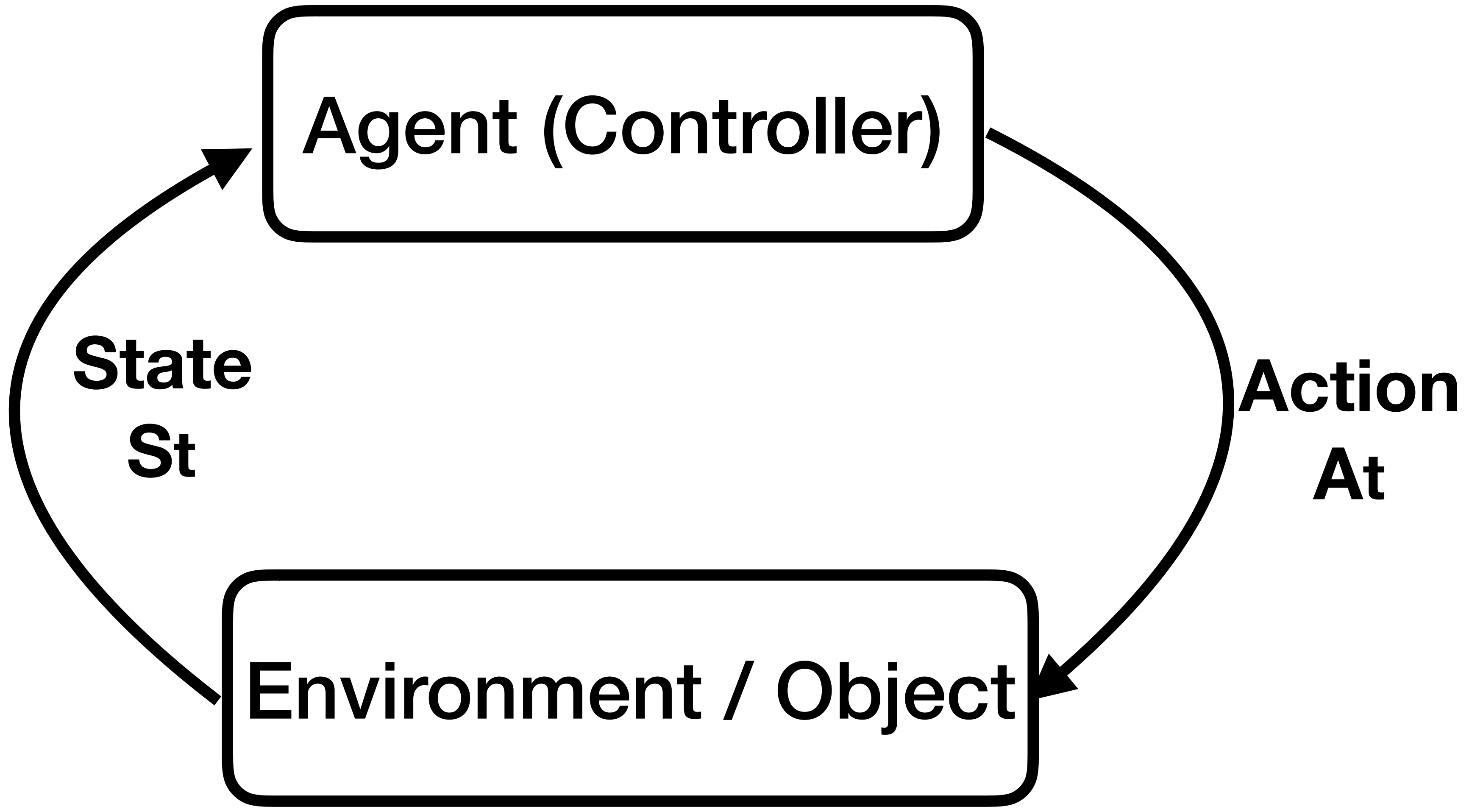


https://en.wikipedia.org/wiki/Q-learning

Reinforcement Learning

[Idea] An agent interacting with an environment, which provides its ***current state*** and numeric ***reward*** signals after each action the agent takes.

[Goal] ***Learn*** how to ***take actions*** in order to ***maximize reward***.

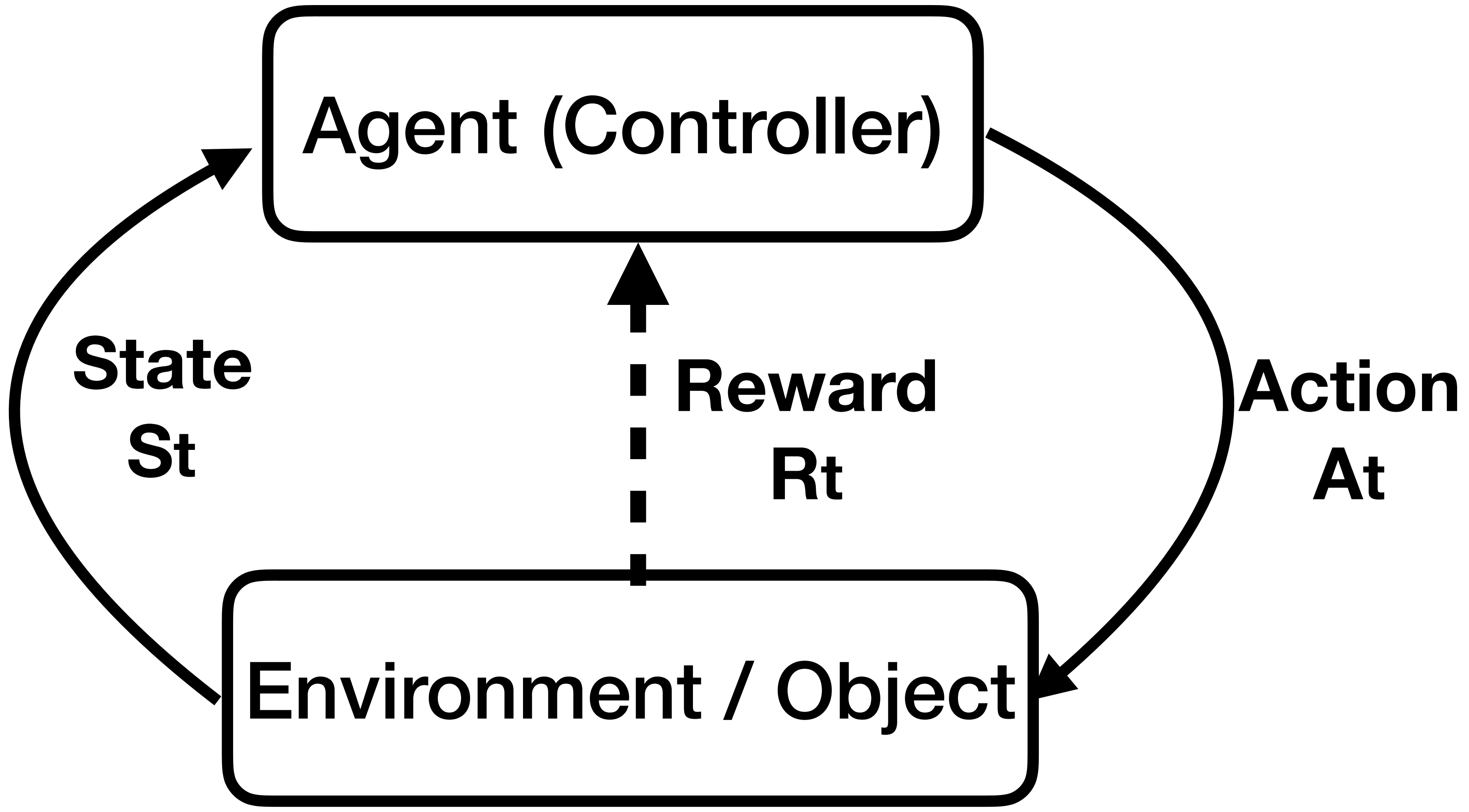


https://en.wikipedia.org/wiki/Q-learning

Reinforcement Learning

[Idea] An agent interacting with an environment, which provides its ***current state*** and numeric ***reward*** signals after each action the agent takes.

[Goal] ***Learn*** how to ***take actions*** in order to ***maximize reward***.

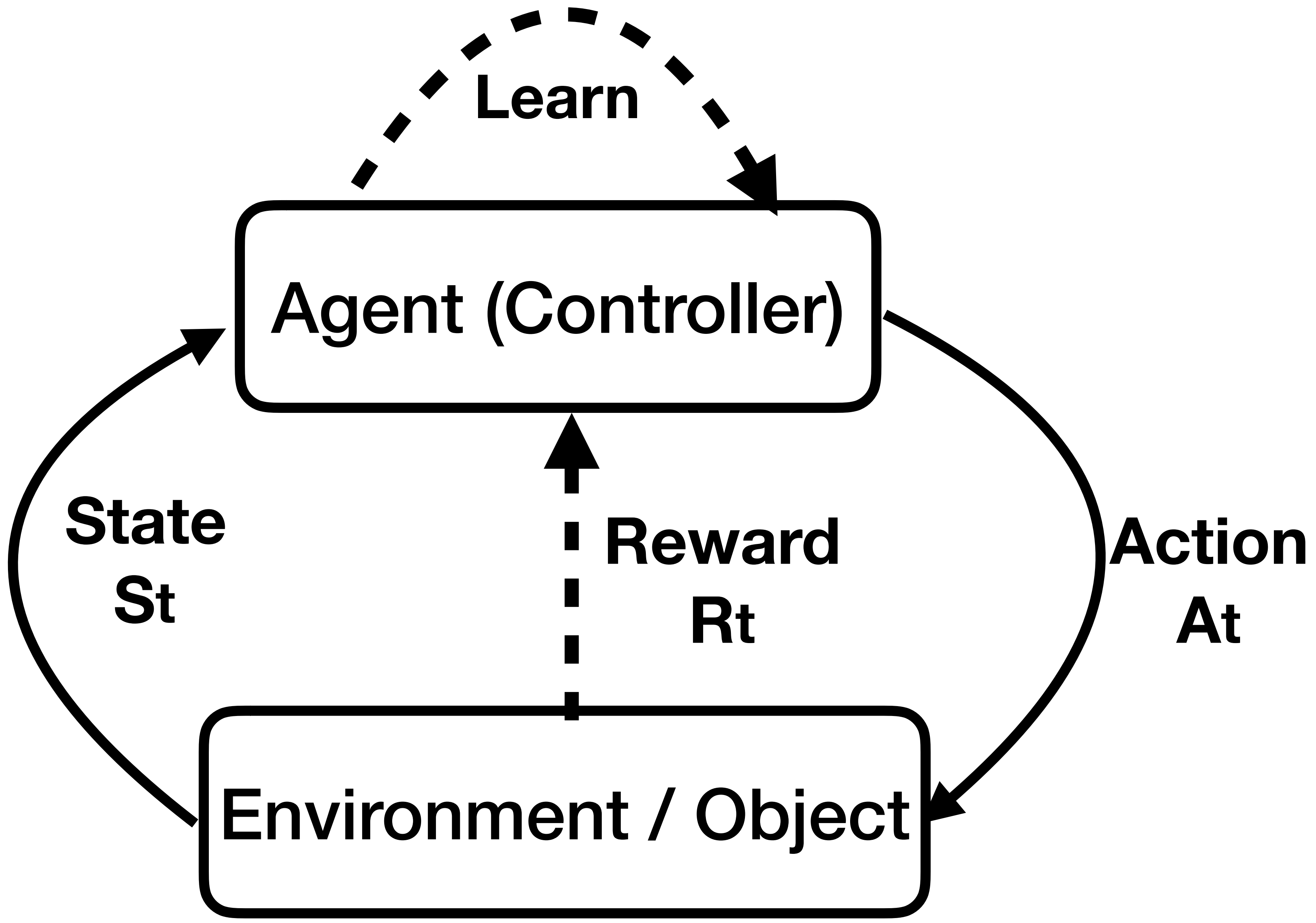


https://en.wikipedia.org/wiki/Q-learning

Reinforcement Learning

[Idea] An agent interacting with an environment, which provides its ***current state*** and numeric ***reward*** signals after each action the agent takes.

[Goal] ***Learn*** how to ***take actions*** in order to ***maximize reward***.



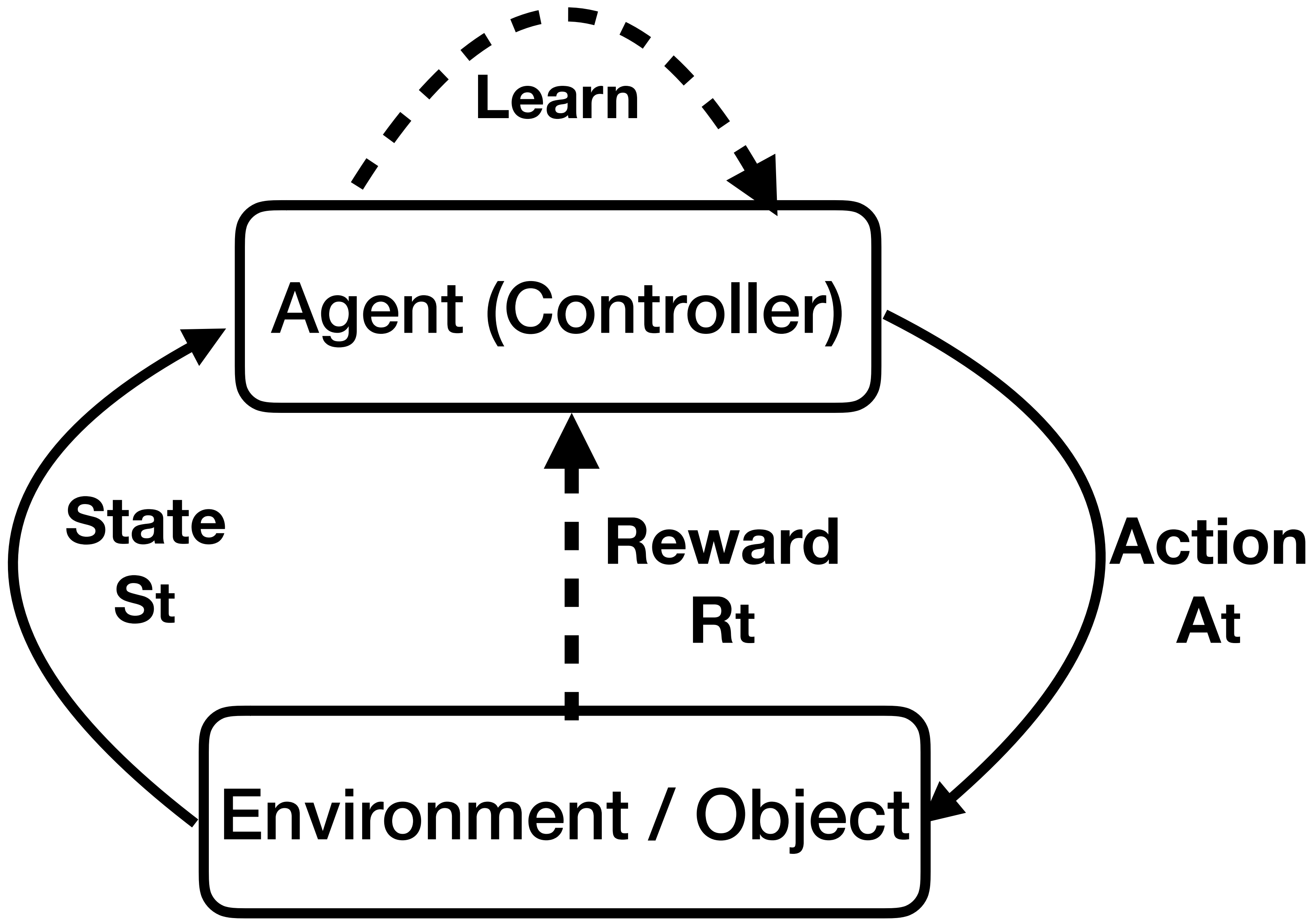
https://en.wikipedia.org/wiki/Q-learning

Reinforcement Learning

[Idea] An agent interacting with an environment, which provides its ***current state*** and numeric ***reward*** signals after each action the agent takes.

[Goal] ***Learn*** how to ***take actions*** in order to ***maximize reward***.

S_t **The state of environment (control object) at any given time t**

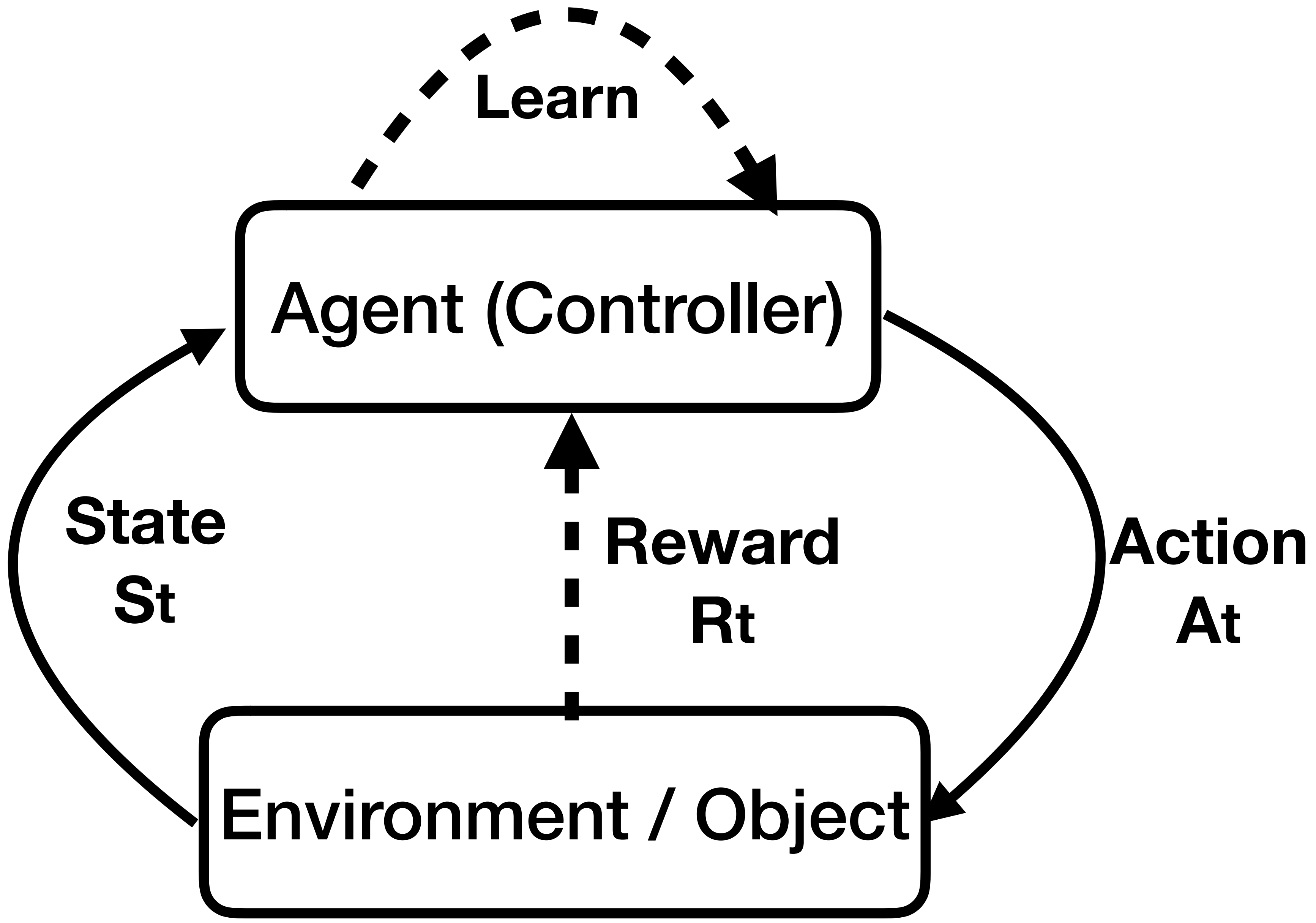


https://en.wikipedia.org/wiki/Q-learning

Reinforcement Learning

[Idea] An agent interacting with an environment, which provides its ***current state*** and numeric ***reward*** signals after each action the agent takes.

[Goal] ***Learn*** how to ***take actions*** in order to ***maximize reward***.



S_t **The state of environment (control object) at any given time t**

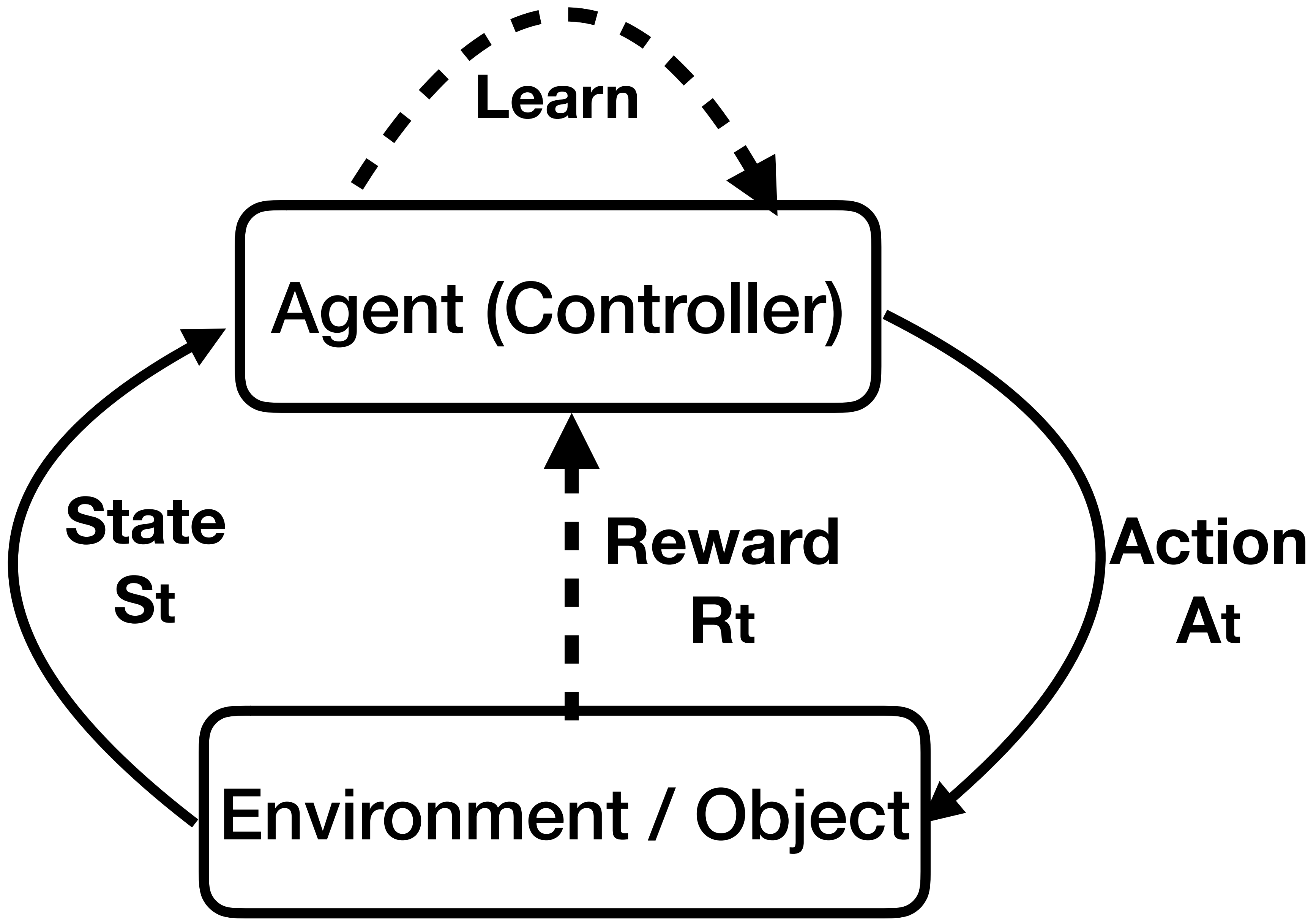
A_t **The corresponding optimal action at any given time t**

https://en.wikipedia.org/wiki/Q-learning

Reinforcement Learning

[Idea] An agent interacting with an environment, which provides its ***current state*** and numeric ***reward*** signals after each action the agent takes.

[Goal] ***Learn*** how to ***take actions*** in order to ***maximize reward***.



S_t **The state of environment (control object) at any given time t**

A_t **The corresponding optimal action at any given time t**

R_t **The actual reward from A_t , i.e., what we want to optimize**

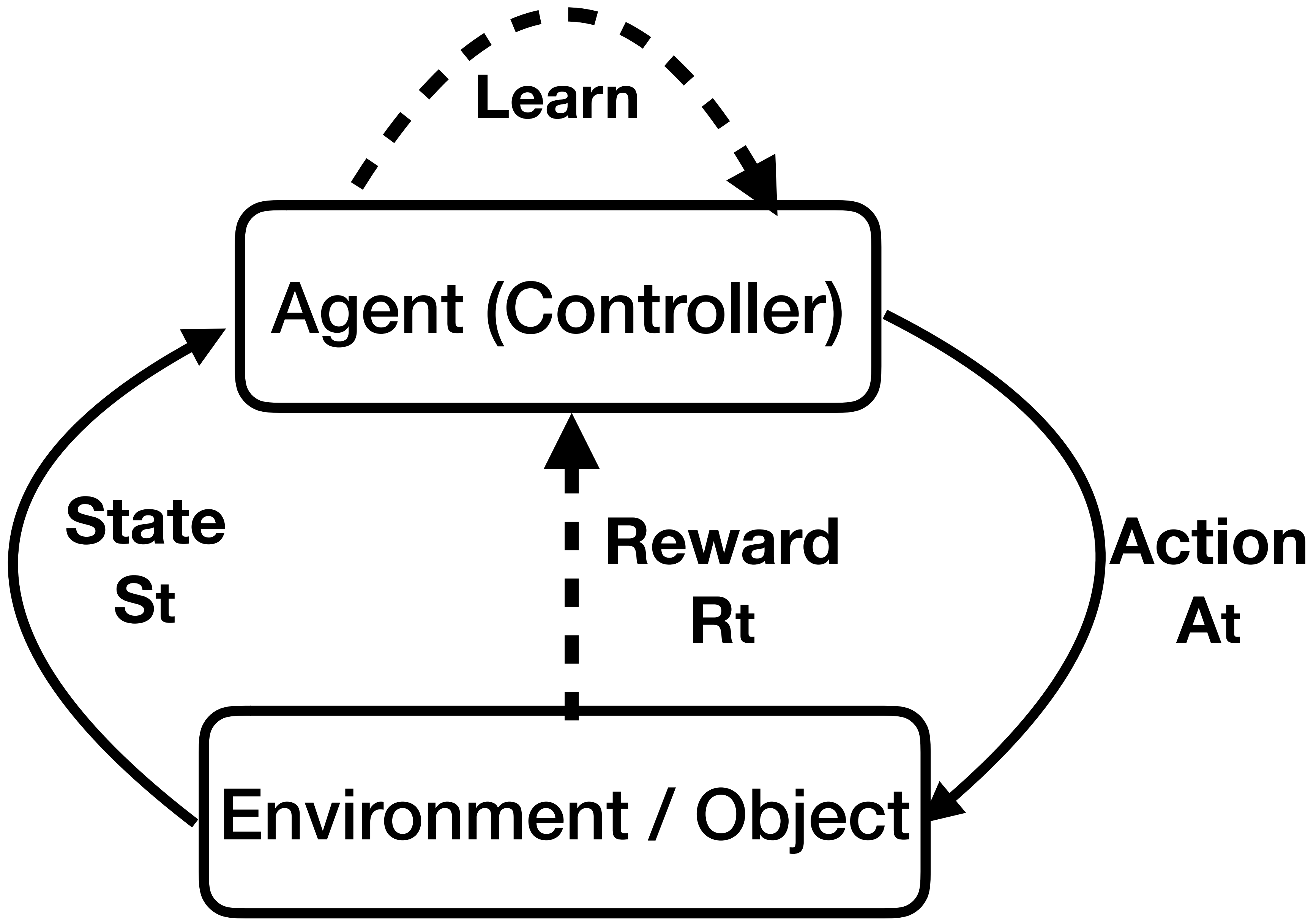
https://en.wikipedia.org/wiki/Q-learning

Argonne NATIONAL LABORATORY

Reinforcement Learning

[Idea] An agent interacting with an environment, which provides its ***current state*** and numeric ***reward*** signals after each action the agent takes.

[Goal] ***Learn*** how to ***take actions*** in order to ***maximize reward***.



S_t **The state of environment (control object) at any given time t**

A_t **The corresponding optimal action at any given time t**

R_t **The actual reward from A_t, i.e., what we want to optimize**

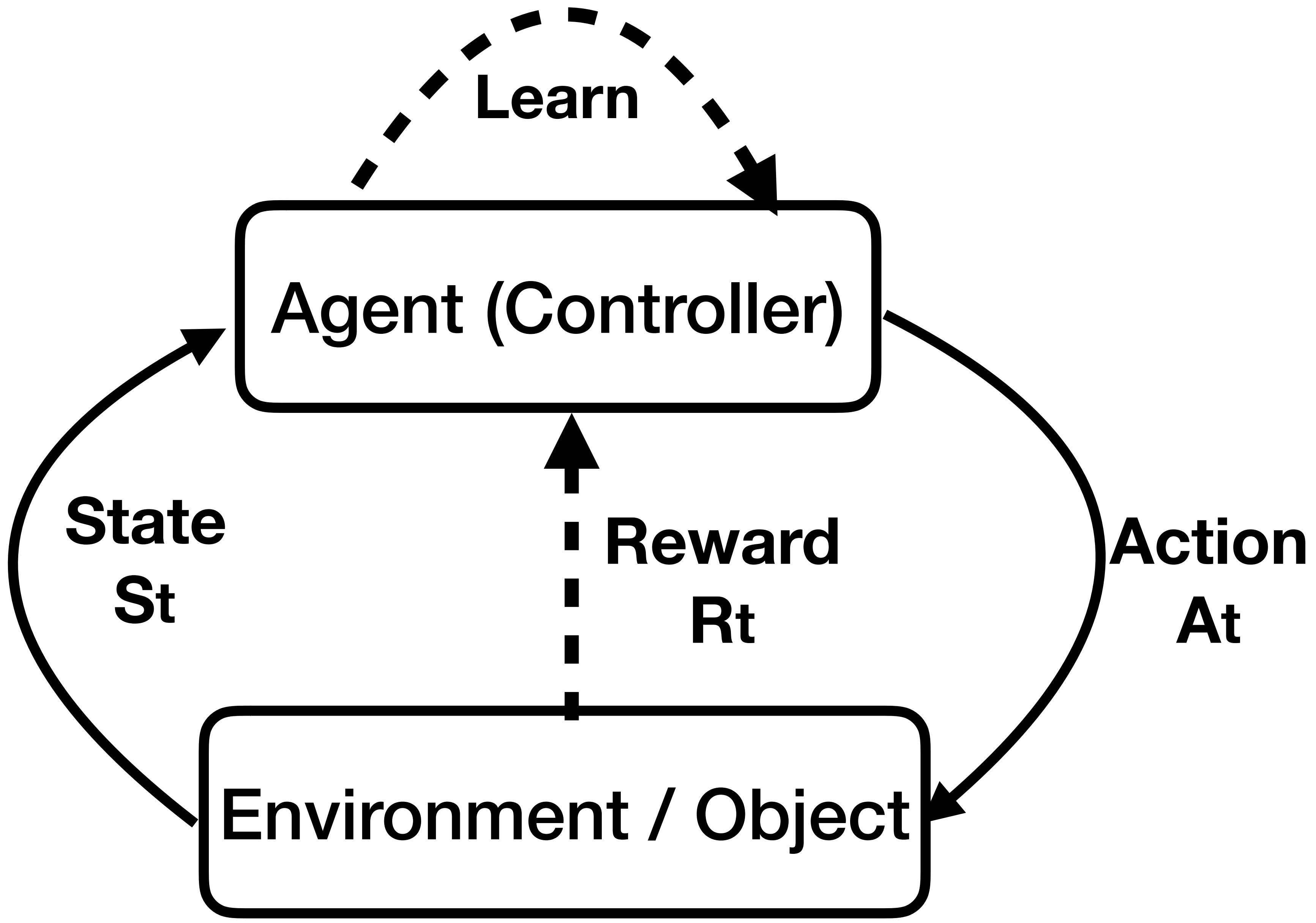
Q-learning

https://en.wikipedia.org/wiki/Q-learning

Reinforcement Learning

[Idea] An agent interacting with an environment, which provides its ***current state*** and numeric ***reward*** signals after each action the agent takes.

[Goal] ***Learn*** how to ***take actions*** in order to ***maximize reward***.



S_t **The state of environment (control object) at any given time t**

A_t **The corresponding optimal action at any given time t**

R_t **The actual reward from** A_t **, i.e., what we want to optimize**

Q-learning

$$Q(s_t, a_t) \leftarrow (1-\alpha) \cdot \underbrace{Q(s_t, a_t)}_{\text{old value}} + \underbrace{\alpha}_{\text{learning rate}} \cdot \left(\overbrace{\underbrace{r_t}_{\text{reward}} + \underbrace{\gamma}_{\text{discount factor}} \cdot \underbrace{\max_a Q(s_{t+1}, a)}_{\text{estimate of optimal future value}}}^{\text{learned value}} \right),$$

https://en.wikipedia.org/wiki/Q-learning

Reinforcement Learning

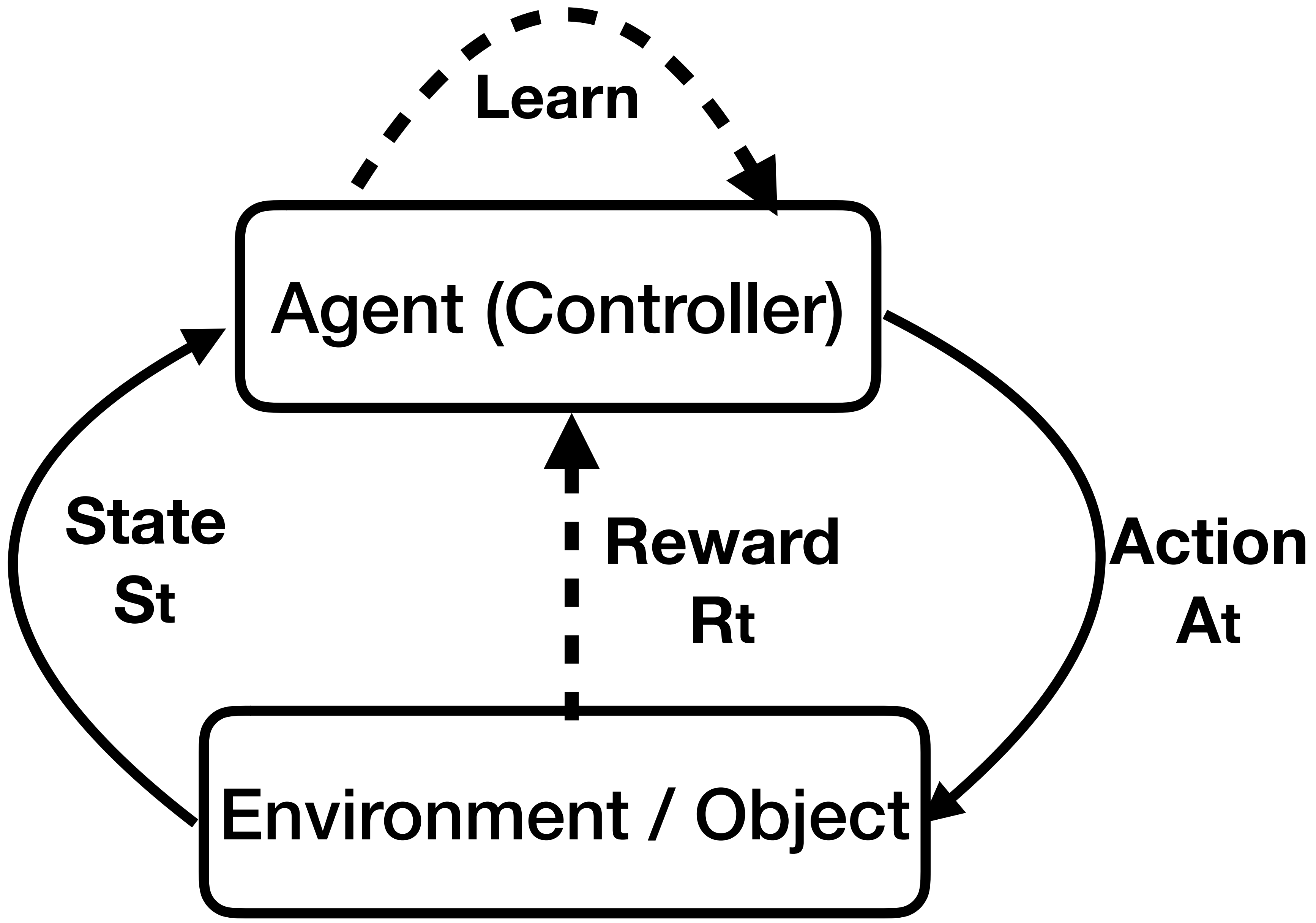
[Idea] An agent interacting with an environment, which provides its ***current state*** and numeric ***reward*** signals after each action the agent takes.

[Goal] ***Learn*** how to ***take actions*** in order to ***maximize reward***.

S_t **The state of environment (control object) at any given time t**

A_t **The corresponding optimal action at any given time t**

R_t **The actual reward from A_t , i.e., what we want to optimize**



Q-learning

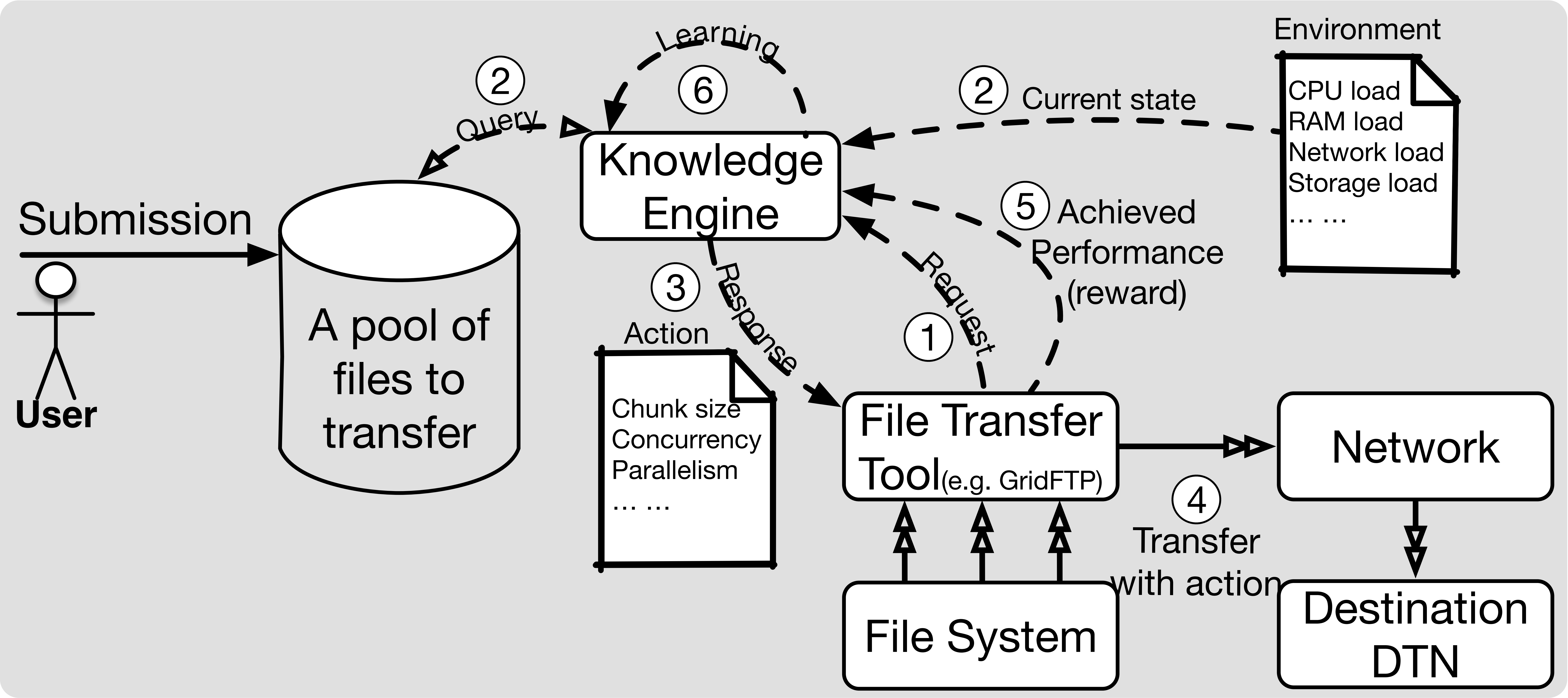
$$Q(s_t,a_t) \leftarrow (1-\alpha)\cdot \underbrace{Q(s_t,a_t)}_{\text{old value}} + \underbrace{\alpha}_{\text{learning rate}} \cdot \left(\overbrace{\underbrace{r_t}_{\text{reward}} + \underbrace{\gamma}_{\text{discount factor}} \cdot \underbrace{\max_a Q(s_{t+1},a)}_{\text{estimate of optimal future value}}}^{\text{learned value}} \right),$$

Policy Gradient

https://en.wikipedia.org/wiki/Q-learning

Smart Data Transfer Node

Workflow

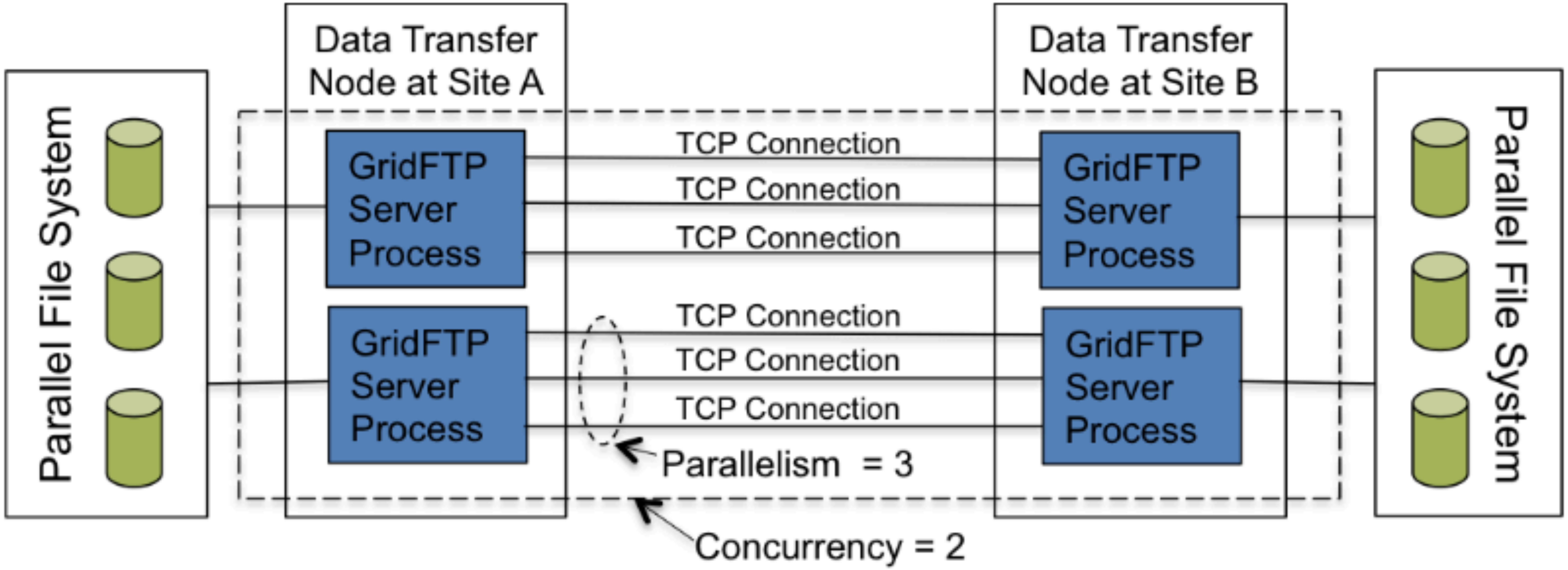


① A file transfer tool requests a file to transfer from the KE. The KE ② checks the current DTN state and ③ responds to the transfer tool with a chunk of file and corresponding optimal transfer parameters (the steering action). ④ The transfer tool transfers the associated chunk with the parameters and monitors the aggregate DTN throughout during this transfer. ⑤ Once completed, DTN's average aggregate throughput is reported to the KE as a reward for its actions. ⑥ Based on the reward (encourage or discourage), the KE updates its internal model parameters to improve its decision policy.

Argonne NATIONAL LABORATORY

State, Action and Reward

Context — High performance wide area data transfer scheme



Argonne NATIONAL LABORATORY

State, Action and Reward

State S_t

- ◉CPU usage (# of GridFTP instance here);
- ◉Total number of TCP streams on DTN;
- ◉The aggregate ingress and egress throughout of the DTN's network interface card;
- ◉The aggregate disk read and write throughput.

Action A_t

- ◉Wether start transferring a new file chunk (True/False). It controls the **total Concurrency**.
- ◉Parallelism used to transfer the file chunk
- ◉The size of file chunk to transfer. It controls the transfer duration, e.g., command frequency.

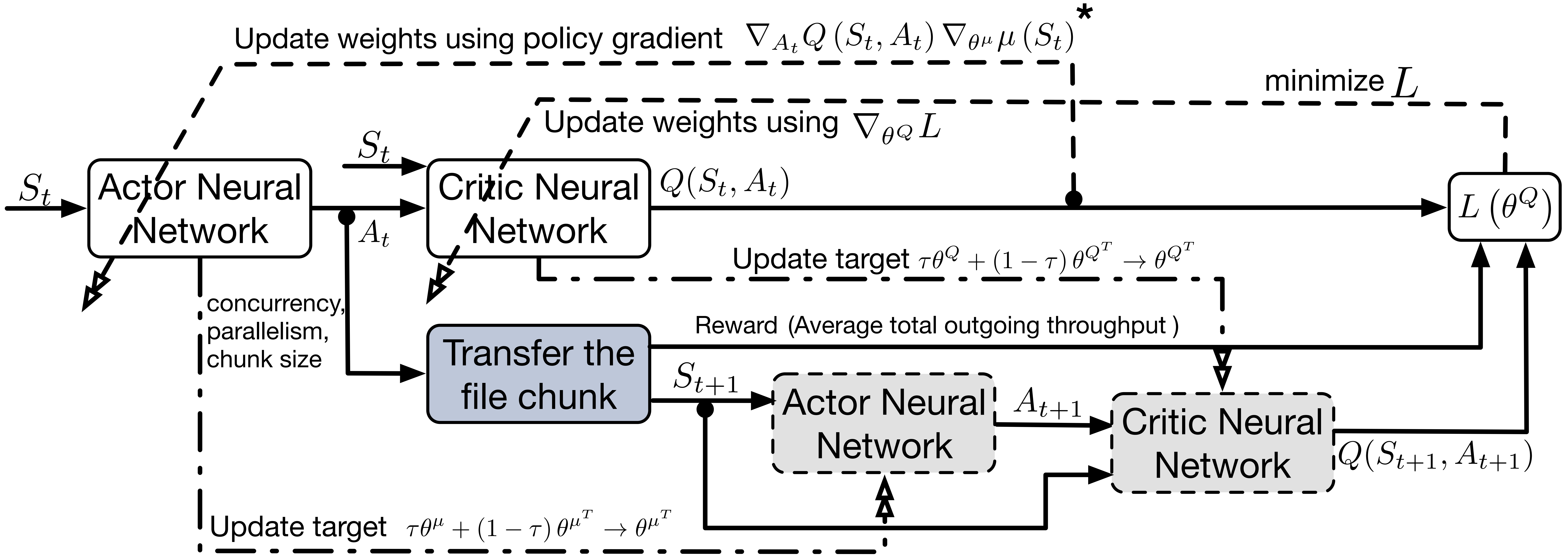
Reward R_t

- ◉The aggregated transfer throughput (of all transfers).

Argonne NATIONAL LABORATORY

Knowledge Engine

Reinforcement learning model architecture



$$y_t = r(S_t, A_t) + \gamma Q(S_{t+1}, A_{t+1})$$

$$L(\theta^Q) = \frac{1}{N}\sum_t [Q(S_t, A_t) - y_t]^2$$

$$\nabla_{\theta^\mu} J \approx \frac{1}{N}\sum_t \nabla_{A_t} Q(S_t, A_t) \nabla_{\theta^\mu} \mu(S_t)^*$$

* D. Silver et al. Deterministic Policy Gradient Algorithms. ICML'14

Argonne NATIONAL LABORATORY

Results and discussion

Reinforcement learning model accuracy versus DTN's aggregated throughput (credit) in dedicated environment.

Effectiveness of the knowledge engine (KE) in a dedicated environment. DTN performance increases as the KE's prediction accuracy improves. (64 iterations per epoch)

Results and discussion

Reinforcement learning model accuracy versus DTN's aggregated throughput (credit) in dedicated environment.

Effectiveness of the knowledge engine (KE) in a dedicated environment. DTN performance increases as the KE's prediction accuracy improves. (64 iterations per epoch)

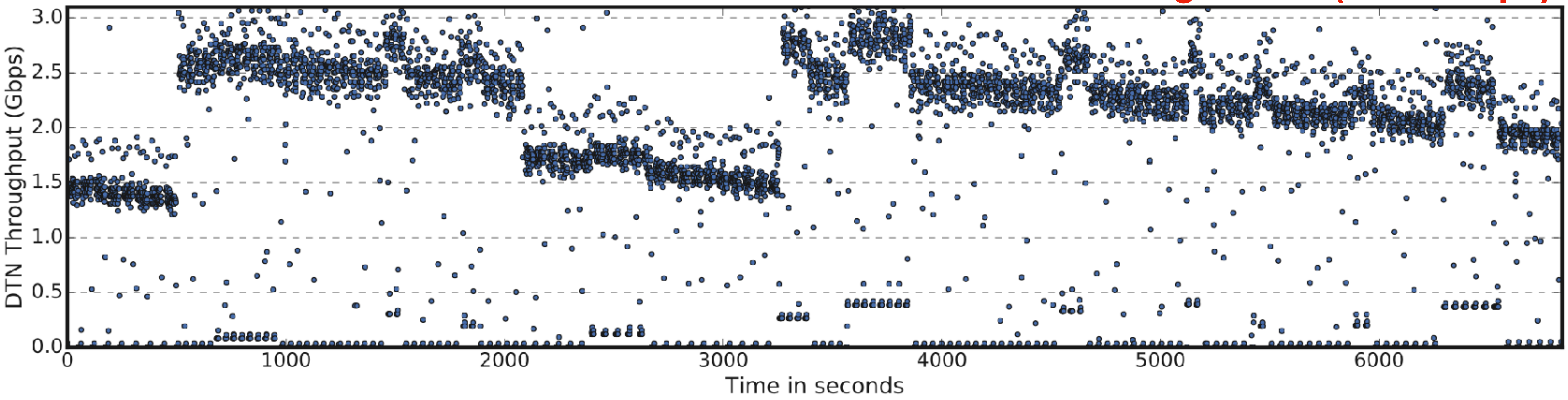
It works!

The knowledge engine is able to find the optimal operating point and, keep DTN working in the optimal operating region.

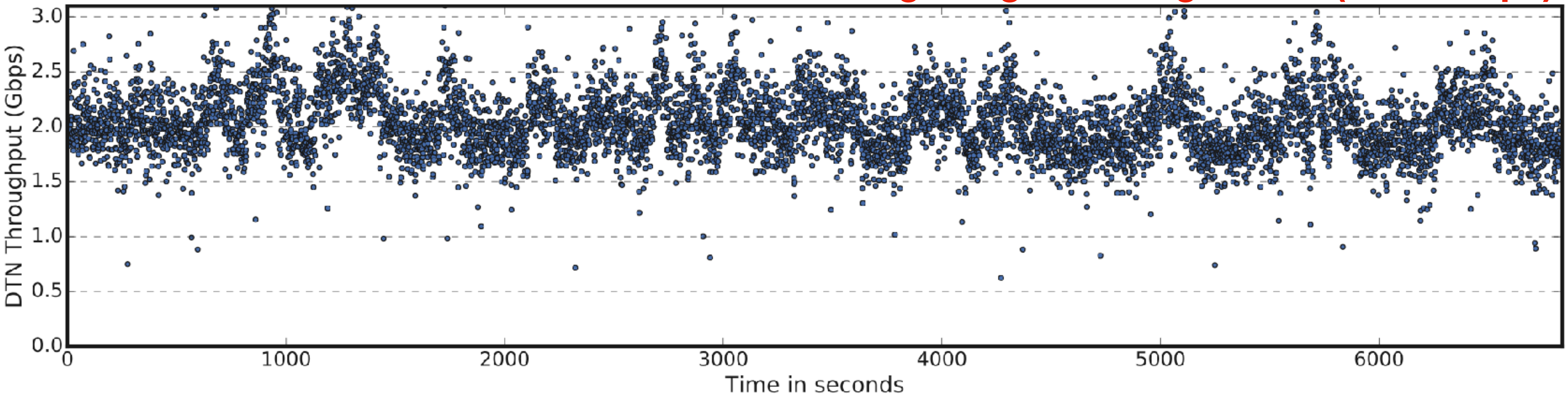
Results and discussion

Experiment in shared environment (adding artificial, reproducible external load to storage)

Heuristic configuration (2.040 Gbps)



Knowledge Engine configuration (2.043 Gbps)



Results and discussion

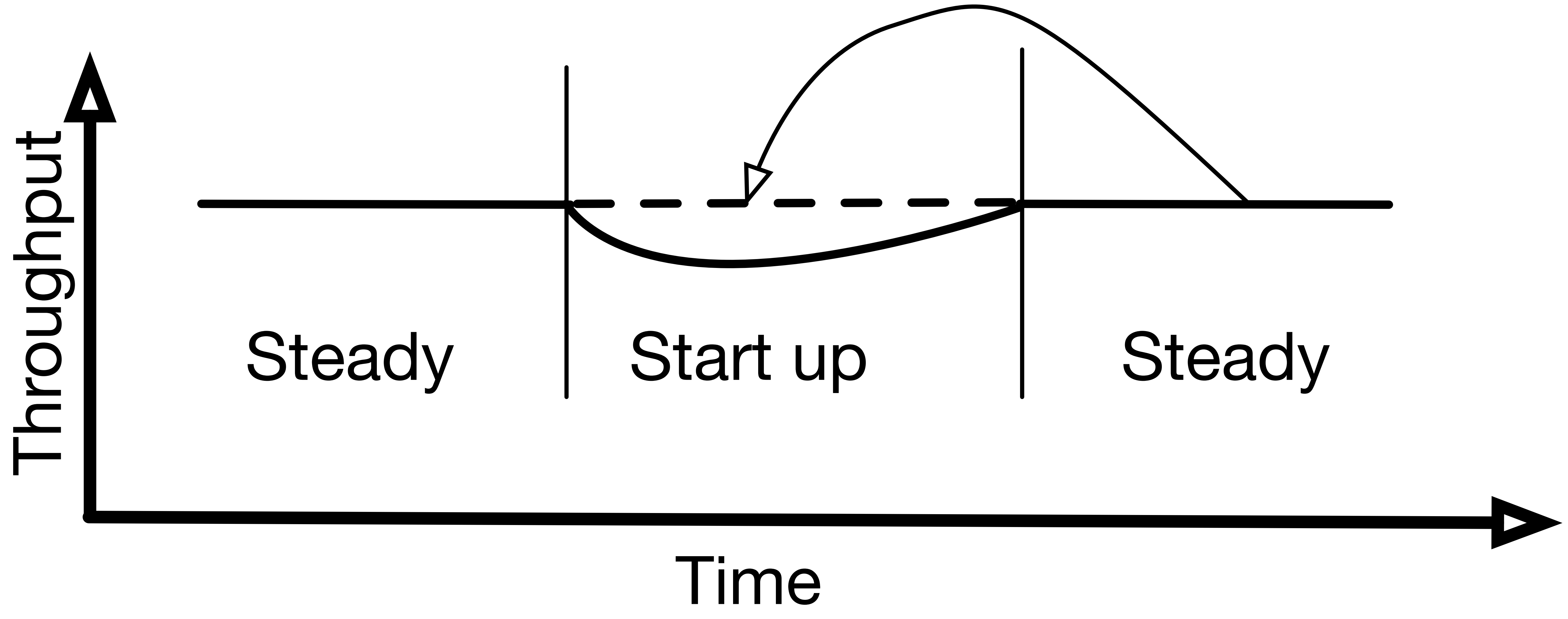
Overhead issue

- GridFTP does not support dynamic concurrency and parallelism.
- We have to restart GridFTP to apply the new parameters.
- There is an overhead for changing parameters.

Results and discussion

Overhead issue

- GridFTP does not support dynamic concurrency and parallelism.
- We have to restart GridFTP to apply the new parameters.
- There is an overhead for changing parameters.

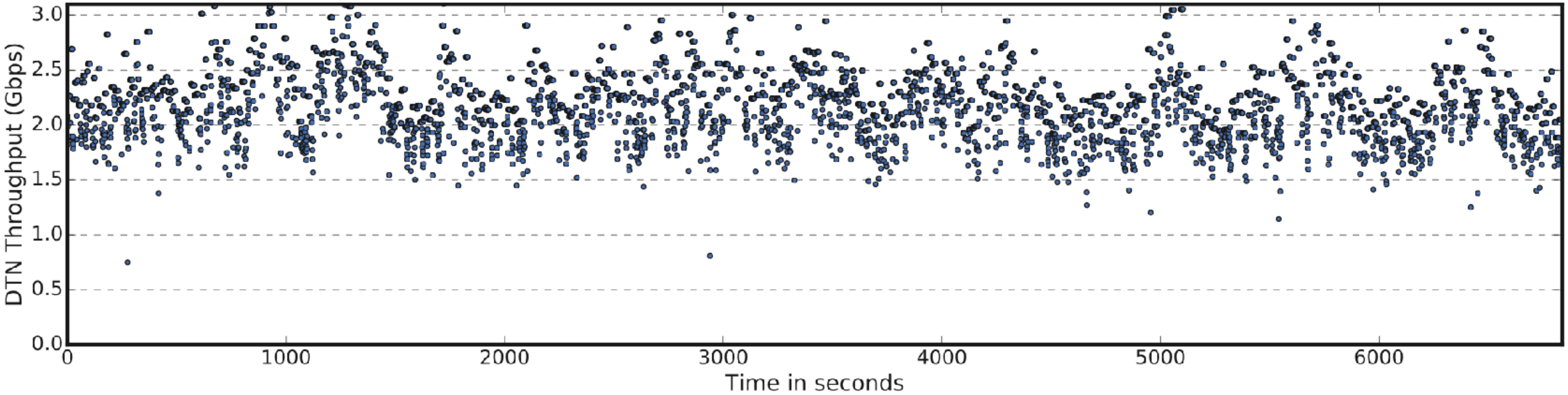


Results and discussion

Overhead issue

- GridFTP does not support dynamic concurrency and parallelism.
- We have to restart GridFTP to apply the new parameters.
- There is an overhead for changing parameters.

Knowledge engine configuration, adjusted to remove overheads (2.273 Gbps)

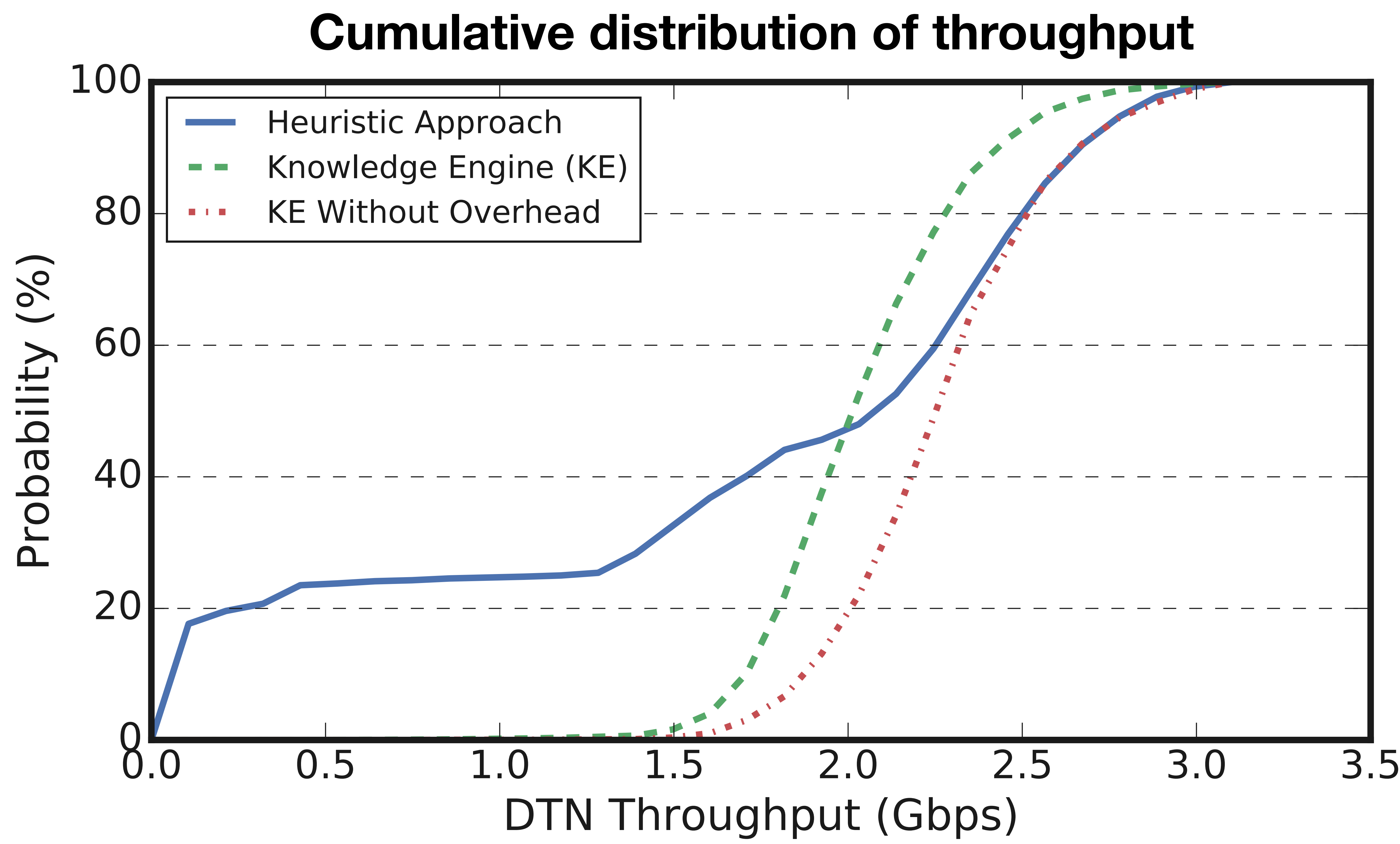


With knowledge engine, we get *about* 11.3% improvement compare with heuristic configuration.

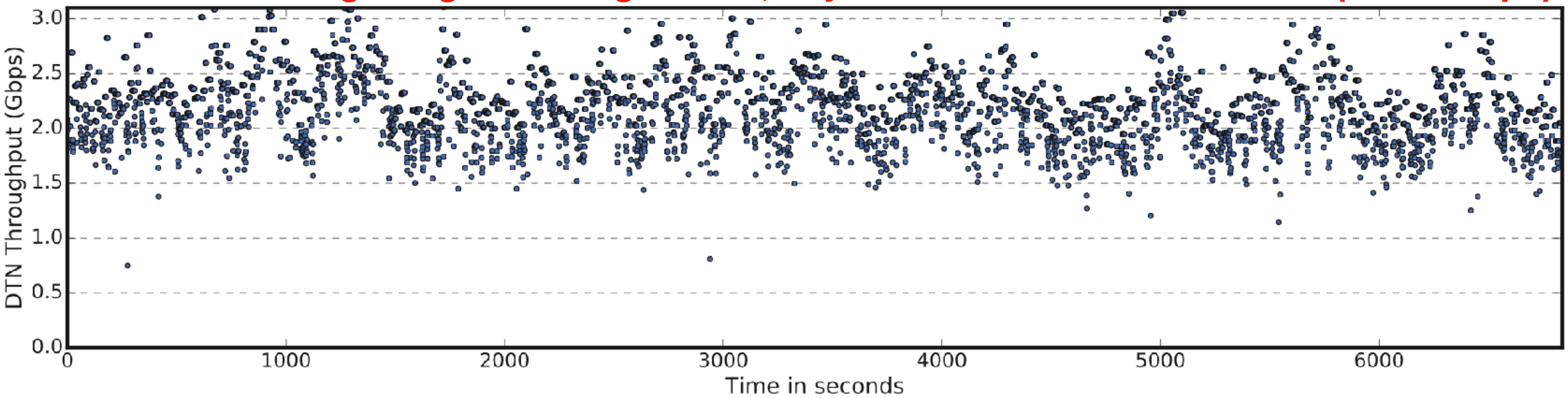
Results and discussion

Overhead issue

- GridFTP does not support dynamic concurrency and parallelism.
- We have to restart GridFTP to apply the new parameters.
- There is an overhead for changing parameters.



Knowledge engine configuration, adjusted to remove overheads (2.273 Gbps)



With knowledge engine, we get *about* 11.3% improvement compare with heuristic configuration.

Conclusion

The knowledge engine that powers the conventional data transfer node with smartness are:

- ☑ **Fully unsupervised, does not need labeled historical data;**
- ☑ **Changes parameters automatically according the state of environment;**
- ☑ **Training is online, self-optimization;**
- ☑ **Suitable for any deployment without specialist;**

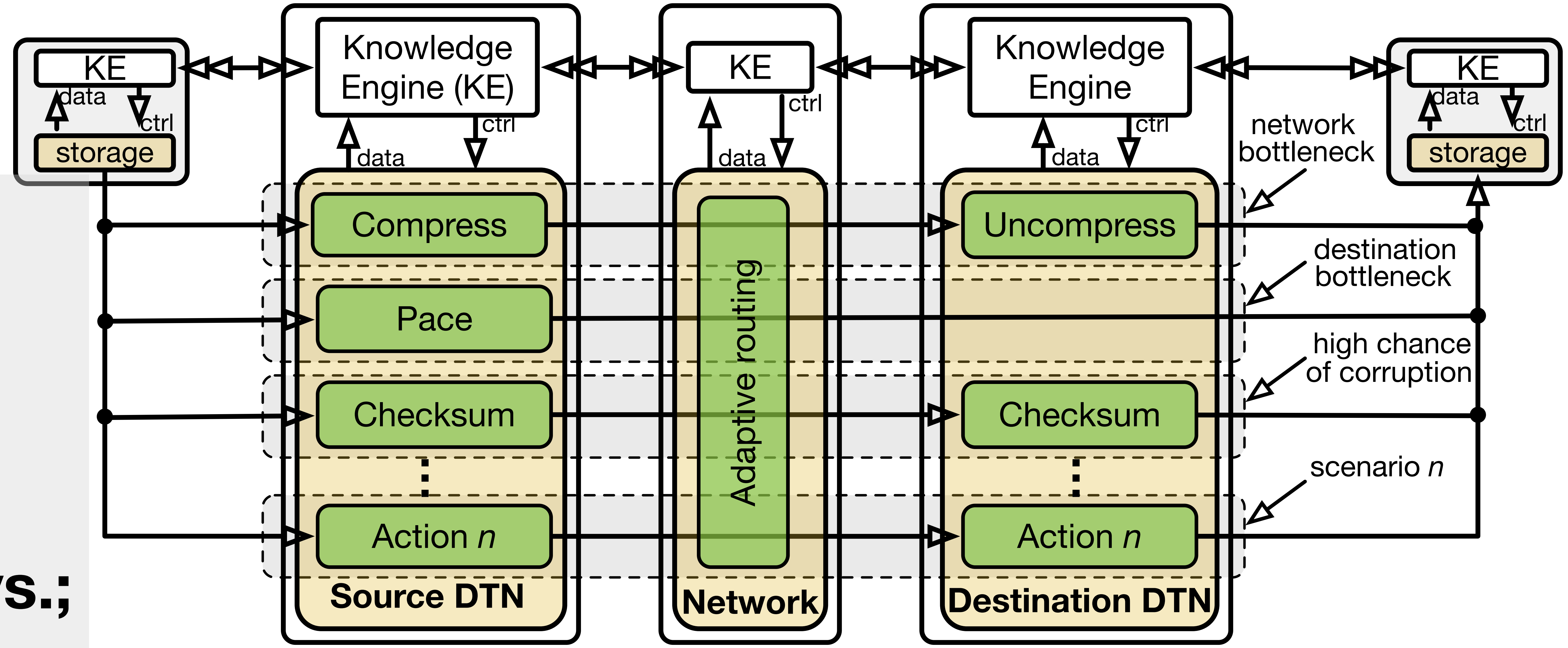
Future work

Future work

- ☑ Tuning more parameters;
- ☑ Testing in practical environment.
- ☑ Embed in distributed workflow;
- ☑ Smart autonomous science ecosys.;

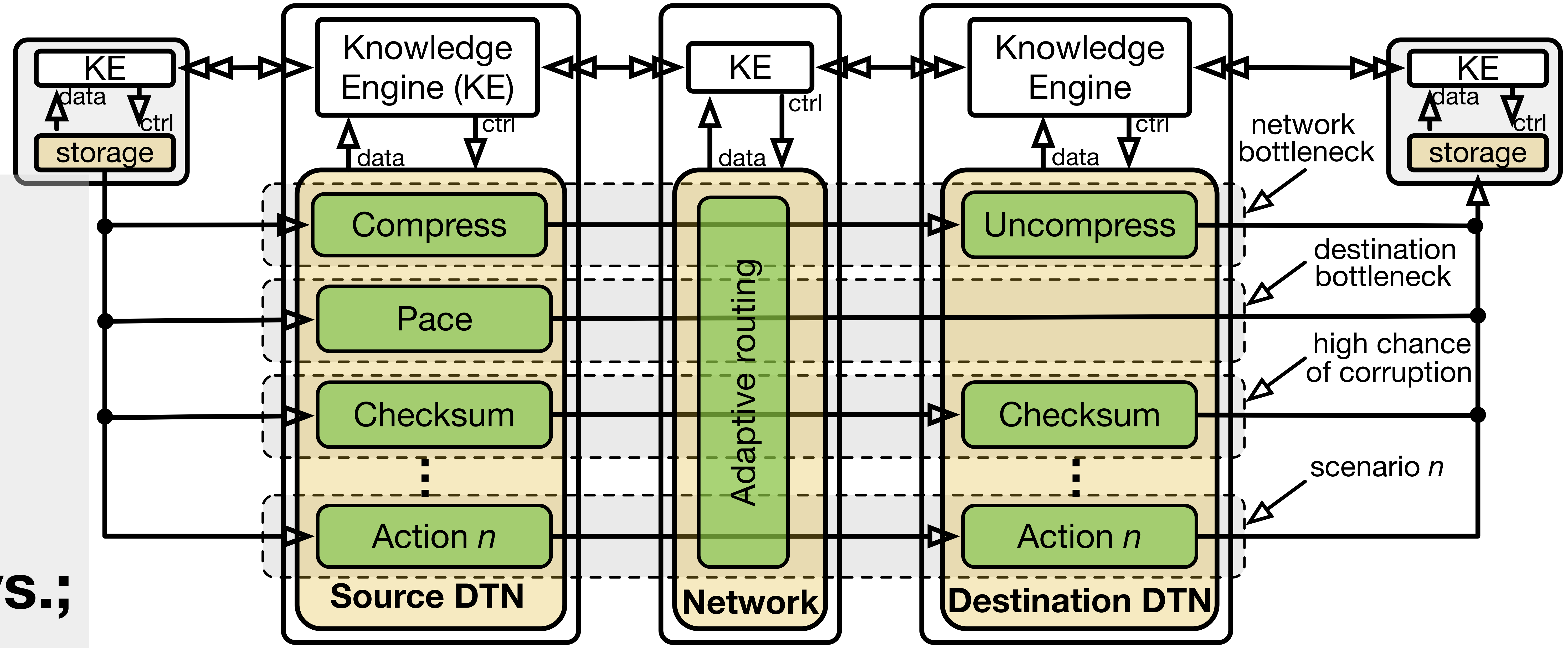
Future work

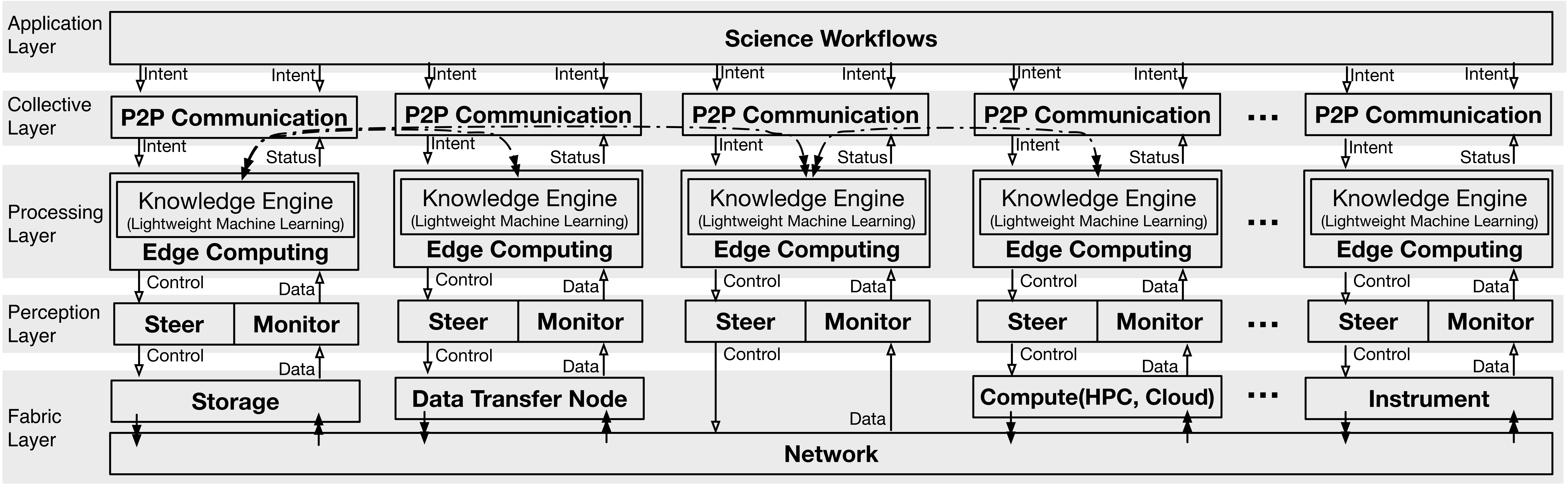
- ☑ Tuning more parameters;
- ☑ Testing in practical environment.
- ☑ Embed in distributed workflow;
- ☑ Smart autonomous science ecosys.;



Future work

- ☑ **Tuning more parameters;**
- ☑ **Testing in practical environment.**
- ☑ **Embed in distributed workflow;**
- ☑ **Smart autonomous science ecosys.;**





Thank you for your attention!

We also want to THANK:

- U.S. Department of Energy, Office of Science, ASCR, and the program manager *Richard Carlson*;
- The Joint Laboratory for System Evaluation (*JLSE*) at Argonne National Laboratory and the *Chameleon* project <www.chameleoncloud.org> for providing resources for testbed.

Q & A

Argonne NATIONAL LABORATORY