Titles

Yaozu Dong[1, 2], Mochi Xue[1,2], Xiao Zheng[2], Jiajun Wang[1,2], Zhengwei Qi[1], Haibing Guan[1]
{*eddie.dong, xiao.zheng*}*@intel.com* {*xuemochi, jiajunwang, qizhenwei, hbguan*}*@sjtu.edu.cn*
[1]*Shanghai Jiao Tong University,* [2]*Intel Corporation*

Abstract

Recently web applications have grown explosively with the widespread of internet and networking devices. Many applications rely on container-base cluster to provide stability and high performance. For a specific application, given thought to diverse resource consumption (such as CPU, memory, etc.) and frequency of different kinds of requests, it is difficult to balance load of servers. So problems like performance degradation will be caused ineluctably if too many requests are sent to over-load servers. In this paper, a dynamic load balance algorithm which takes request consumption and server load into account is proposed to reduce the performance problem. Experimental results are given to validate the performance of proposed algorithm.

1 Introduction

The rapid growth of internet introduces a new economy pattern named internet economy, under which almost all activities can be supported by web services through PCs or mobile devices. Aiming to achieve high resource utilization, web applications tend to be deployed into cluster based on a lightweight virtualization called container just like other traditional applications do. Unlike traditional applications which only execute job assigned in advance, web services have to listen on users' requests and responses as soon as possible. Workload of web services are changing unpredictably because they offer varied businesses, each business consists of operations that may consume CPU, allocate memory or access disk, and it is impossible to know which business will be executed in any time. The unpredictable workload introduces a hug challenge on load-balancing. Unfortunately, the container-based solutions like Kubernetes fail to cope with this challenge, their fair load-balancing solution performs well only when workload are the same, which is contrary to workload of web services, as a result, web services fail to gain high resource utilization, it loses stability and performance conversely.

To address the load balance problem, this paper introduce a dynamic load-balancing solution based on Kubernetes, a popular container cluster management system. The new solution mainly covers two issues: Load collection and load-balancing algorithm. To address the load collection issue, we implement a load monitor to evaluate resource consumption of requests and collect load of backend containers. Through collecting and analyzing logs of web services, monitor generates description for different request in requirement of CPU, memory, disk and network. Through listening on status changes of container, monitor generates load description for usage of CPU, memory and network.

For the algorithm part, we introduce request and container models based on description generated by monitor. The generation of models considers not only the load description, but also other details such as the locality, deployment configuration of container and so on. Request dispatching is done through matching request models to container models.

In summary, this paper makes the following contributions:

- A load monitor for learning resource consumption of requests and listening load changes of containers.
- A dynamic load-balancing algorithm which combines resource consumption of request and real-time load of container to dispatch request and finally achieve stability, high performance and high resource utilization for web services.
- An evaluation to show the performance of dynamic load-balancing solution.

The rest of the paper is organized as follows: Section 2 describe background information about performance of web services in container-base cluster. Section 3

presents the design and implementation of dynamic load-balancing solution. Then section 4 evaluate the solution. Section 5 shows the related work, and finally section 6 discuss on the future work.

The rest of the paper is organized as follows: Section 2 describes some background information on gVirt and Graphic Process Unit (GPU) programming model. Section **??** presents our benchmark for media transcoding and discusses the Massive Update Issue in detail, followed by the design and implementation of gHyvi In section **??**. Then, section 4 evaluates the gHyvi and section 5 discusses the related work. Finally, section 6 concludes with a brief discussion on future work.

2 Background

Performance of web services is sensitive to system metrics of their backend servers, in container-based cluster, that is backend containers. For convenience, we called the backend containers as endpoints in this paper. This section first introduces the current load-balancing solution in Kubernetes, and then points out factors that affect performance of web services in container-based cluster.

2.1 Ingress controller

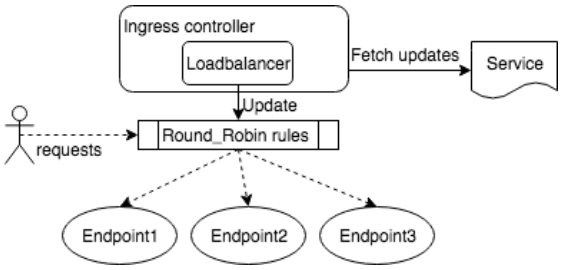

Figure 1: Workflow of ingress controller

Figure 1 shows the workflow of kubernetes load-balancing component – ingress controller: To balance network workload, ingress controller focuses on two part, one is to keep track of the IP and number change of endpoints so that it can update traffic rules in time to guarantee accuracy. The other part is watching ingress to update load-balancing strategy. Ingress is a resource describing inbound rules for connections to reach endpoints. With correct status of endpoints and a configured ingress, ingress controller sets up appropriate traffic rules for its built-in loadbalancer(nginx). Finally, the loadbalancer dispatches requests to suitable endpoints according to configured rules, balancing load across endpoints.

2.2 Influencing factors

Factors that affect the performance include CPU usage, memory usage, network usage, sharing problem and locality. Papers have demonstrated that servers with high usage in CPU, memory or network will perform badly, and as a result, it causes delay on requests processing. Things are same when it comes to endpoints. Thus, if we demand high performance for web application, we?d better even load of each pod and reduce overhead endpoints as many as possible. Sharing problem occurs when several endpoints share one database. Traditionally, developers tend to reduce pressure of database by dividing it into several databases and synchronize data between them. Therefore, it is common for an endpoint to access a database that is not only the same as some endpoints, but also different from others. Under this circumstance, performance of web applications can be improved through reducing race between endpoints that share the same database. Locality means that the placement of endpoints may affect their performance. Though containers are claimed to isolate resource from each other, papers still show that when two or more endpoints are placed in the same node, the delay appears. More and more, although under circumstance that an endpoint monopolizes a node, it still remains possibility for delay when the node is in high-load status because of some system processes. So the locality should also be considered for a better load-balancing performance.

2.3 Shortcoming of ingress controller

Rather than balancing load, ingress controller concentrates more on linking ingress to nginx features like session affinity to let users to custom their load-balancing algorithm smoothly in kubernetes. However, its configurable algorithms are simply based on static algorithms like round_robin, which do not consider influencing factors mentioned above. To address this, this paper presents an alternative basic algorithm to ingress controller, that is a dynamic load algorithm that takes these factors and request information into consideration.

3 Design and implementation

To overcome the shortcoming of ingress controller, this paper proposes a dynamic load-balancing solution. The solution consists of two part, one part is load monitor, the other part is dynamic load-balancing algorithm.

3.1 Workflow

Figure 2 illustrates the workflow of the dynamic load-balancing solution:

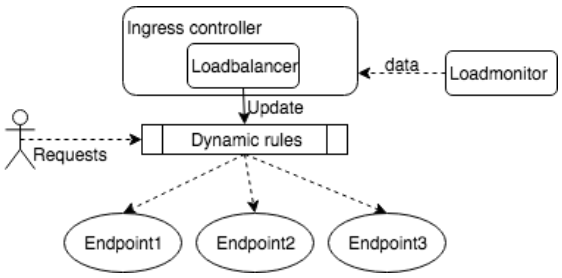

Figure 2: Workflow of dynamic solution

(1) Load monitor gather necessary data for dynamic load balancing and offer data to loadbalancer.

(2) Loadbalancer makes a further evaluation for data received from loadmonitor and dispatches incoming requests based on dynamic load-balancing algorithm.

3.2 Load monitor

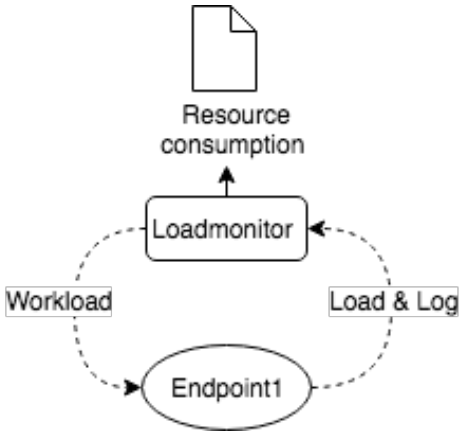

Figure 3: request-learing phase

Load monitor collects, analyzes and outputs information needed by dynamic load-balancing algorithm. Load monitor processes two kinds of information: request logs and endpoint loads. Request logs are used for evaluating resource consumption of request. Before running in kubernetes, each web service goes through the request-learning phase shown by figure 3: First, a single endpoint is deployed, then load monitor generates workload by keep on sending requests in a fixed speed to the endpoint, the speed increases with time until it reach the processing limit of endpoint. At the same time, load monitor combines service logs, workload and endpoint load to evaluate resource description for each type of request. We define RD_i as resource description of a i_{st} type request.C_i ,M_i and N_i stand for CPU, memory, network consumption to process a i_{st} type request. S_i^t stands for speed of sending i_{st} type request in timestamp t. c_i^t, m_i^t, n_i^t are CPU, memory, network load of endpoint in timestamp t during i_{st}request-learning phase. The computation of resource description is as follow:

$$C_i = \frac{\sum_{j=1}^{k} \frac{c_i^t}{s_i^t}}{k}, M_i = \frac{\sum_{j=1}^{k} \frac{m_i^t}{s_i^t}}{k}, N_i = \frac{\sum_{j=1}^{k} \frac{n_i^t}{s_i^t}}{k}$$

$$RD_i = \{C_i, M_i, N_i\}$$

Then load monitor enters load-collecting phase shown in figure 4 to generate load description for endpoints.To increase load-balancing performance, load description of endpoints has to cover all influencing factors. For the resource usage part, Load monitor collects system loads from kubelet, a component that is not only responsible for managing lifecycle of pods, but also takes charge of collecting cluster load. For sharing problem and locality, load monitor obtains endpoint information from apiserver. Endpoint information includes configuration that which endpoints share the same database, which endpoints are deployed in the same node and so on. Table 1 is an example of load description given by load monitor.

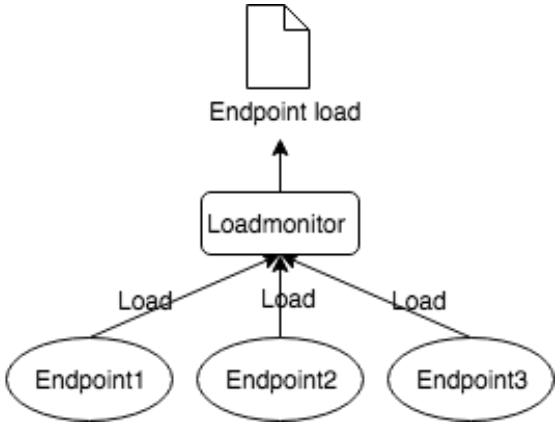

Figure 4: load-collecting phase

3.3 Dynamic load-balancing algorithm

The dynamic load-balancing algorithm considers request consumption and endpoint load to even workload across endpoints. It aims to reduce overload endpoints and promote RPS. The algorithm is divided into two phase: Classification phase and dispatching phase. During classification phase, we evaluate requirement of requests, utilization of nodes and ability of endpoints and classify them. For evaluating requests, we let requests compare

Type	Pod	Pod	Node
Name	App1	Db1	Node1
CPU limit	10	10	50
CPU usage	2	3	20
Memory limit	100M	200M	2G
Memory usage	20M	100M	500M
Network limit	1M/s	2M/s	5M/s
Network usage	500k/s	1M/s	3M/s
Node	Node1	Node1	-
Database limit	Db1	-	-

Table 1: Load description

with each other to get their consumption level of CPU, memory and network, that is to know whether their consumption is high or not compared to others. For evaluating node, we simply adopt utilization rate under the assumption that hardware configuration of all nodes in kubernetes are the same. For the pod part, first we compare usage of CPU, memory and network among pods to get their load level, then we take load level, utilization rate and ability of node and database that pods rely on in to consideration to evaluate ability of endpoints. The evaluation is as follow:

1. Calculate requirement level of each request type in CPU, memory and network:

$$L_{max} = Max\{L\_f_1, ..., L\_f_i\}$$

$$L_{min} = Min\{L\_f_1, ..., L\_f_i\}$$

$$r\_f_i = \frac{L\_f_i - L_{min}}{L_{max} - L_{min}} \cdot 100\%$$

$$R\_f_i = \begin{cases} High & r\_f_i \geq T_r \\ low & r\_f_i < T_r \end{cases}$$

where L_f_i is consumption of a given factor among CPU, memory and network, r_f_i represents consumption of i_{st} request type among all request types.R_f_i is requirement level. For each request type, its requirement is $\{R\_CPU, R\_MEM, R\_NET\}$. By default, we set the threshold $T_r = 0.5$.

2. Calculate utilization of node in CPU, memory and network:

$$U_i = \frac{L\_usage_i}{L\_limit_i} \cdot 100\%$$

where L_usage_i and L_limit_i denote usage and limit of a given factor among CPU, memory and network, U_i represents utilization of i_{st} node.For each node, its utilization is $\{U\_CPU, U\_MEM, U\_NET\}$.

3. Calculate ability of pod in CPU, memory and network:

$$L_{max} = Max\{L\_usage_1, ..., L\_usage_i\}$$

$$L_{min} = Min\{L\_usage_1, ..., L\_usage_i\}$$

$$U_i = \frac{L\_usage_i}{L\_limit_i} \cdot 100\%$$

$$P_i = \frac{L\_usage_i - L_{min}}{L_{max} - L_{min}} \cdot 100\%$$

$$A_i = 1 - (U_i \cdot \alpha + P_i \cdot \beta + N_i \cdot \gamma + D_i \cdot \delta)$$

$$(\alpha + \beta + \gamma + \delta = 1)$$

where L_usage_i denotes usage of a given factor, L_limit_i is the limit, U_i stands for utilization rate, P_i represents load level, N_i and D_i are utilization of node and ability of database that pod relies on. For database pods, the D_i is zero.By default, we set $\alpha = 0.5, \beta = 0.1, \gamma = 0.2, \delta = 0.2$. For each pod, its ability is $\{A\_CPU, A\_MEM, A\_NET\}$.

During dispatching phase, endpoints are sequenced by their ability: Firstly, we set up thresholds to group endpoints. That is $T_c = 70\%$ for deciding whether endpoint's CPU ability is in high or normal level, $T_m = 50\%$ for deciding whether endpoint's memory ability is high or normal, $T_n = 50\%$ for deciding whether endpoint's network ability is high or normal. As a result, endpoints are grouped into 8 groups. Table 2 shows definition of these groups.

Group	CPU	Memory	Network
g1	$A\_CPU \geq T_c$	$A\_MEM \geq T_m$	$A\_NET \geq T_n$
g2	$A\_CPU \geq T_c$	$A\_MEM \geq T_m$	$A\_NET < T_n$
g3	$A\_CPU \geq T_c$	$A\_MEM < T_m$	$A\_NET \geq T_n$
g4	$A\_CPU \geq T_c$	$A\_MEM < T_m$	$A\_NET < T_n$
g5	$A\_CPU < T_c$	$A\_MEM \geq T_m$	$A\_NET \geq T_n$
g6	$A\_CPU < T_c$	$A\_MEM \geq T_m$	$A\_NET < T_n$
g7	$A\_CPU < T_c$	$A\_MEM < T_m$	$A\_NET \geq T_n$
g8	$A\_CPU < T_c$	$A\_MEM < T_m$	$A\_NET < T_n$

Table 2: Endpoint groups

A new endpoint is inserted into a desired group by insertion sort algorithm in descending order first by A_CPU, then A_MEM and finally A_NET. The algorithm 1 describes the insertion.

The dispatch decision is made based on request requirement level and endpoint groups. The steps of dispatching a request is as follow:

Algorithm 1 modified insertion sort algorithm

Require: $g[]$, $endpoint$
1: $i \leftarrow 0$
2: **while** $i < sizeof(g)$ **do**
3: **if** $!(endpoint.A\_CPU \geq g[i].A\_CPU)$**and**
4: $(endpoint.A\_MEM \geq g[i].A\_MEM)$ **then**
5: $i++$
6: continue
7: **else if** $endpoint.A\_CPU \geq g[i].A\_CPU$ **then**
8: break
9: **end if**
10: i++
11: **end while**
12: DO_INSERTION$(g, i, endpoint)$

Algorithm 2 selection algorithm

Require: $hg[\{ep|ep\ \varepsilon\ g1 \cup g2 \cup g3 \cup g5\}]$,
$lg[\{ep|ep\ \varepsilon\ g4 \cup g6 \cup g7 \cup g8\}]$,$request$
1: **function** RANDOM_CHOICE(hg,lg)
2: **if** $sizeof(hg) == 0$ **then**
3: **if** $sizeof(lg) \leq 3$ **then**
4: **return** lg
5: **else**
6: $i, j, k =$ RANDOM_3INT$(0, sizeof(lg))$
7: **return** $\{lg[i], lg[j], lg[k]\}$
8: **end if**
9: **else if** $sizeof(hg) \leq 3$ **then**
10: **return** hg
11: **else**
12: $i, j, k =$ RANDOM_3INT$(0, sizeof(hg))$
13: **return** $\{hg[i], hg[j], hg[k]\}$
14: **end if**
15: **end function**
16: **function** SCORECANDIDATE(candidate,request)
17: $i \leftarrow 0$
18: **while** $i < sizeof(candidate)$ **do**
19: $S\_c \leftarrow candidate(i).A\_CPU * request.\theta$
20: $S\_m \leftarrow candidate(i).A\_MEM * request.\lambda$
21: $S\_n \leftarrow candidate(i).A\_NET * request.\mu$
22: $candidate(i).Score = S\_c + S\_m + S\_n$
23: **end while**
24: SORTBYSCORE$(candidate)$
25: **return** $candidate$
26: **end function**
27: $c \leftarrow$ RANDOM_CHOICE(hg, lg)
28: $c \leftarrow$ SCORECANDIDATE(c)
29: **if** $request.priority == High$ **then**
30: DISPATCH_REQUEST$(request, highestof(c))$
31: **else**
32: DISPATCH_REQUEST$(request, middleof(c))$
33: **end if**

1. Query request type and corresponding requirement level.

2. Select suitable endpoint for request, the selection is decided by selection algorithm 2. First of all, we define priority of endpoint groups: we regard g4,g6,g7,g8 as low-priority groups and g1,g2,g3,g5 as high-priority groups because of the fact that endpoints have poor performance when they score badly in more than one factor. Then we pick three endpoints as candidates randomly from high-priority groups, grade them and choose the most suitable one. The equation to grade endpoints is as follow:

$$G_i = A\_CPU_i \cdot \theta + A\_MEM_i \cdot \lambda + A\_NET \cdot \mu$$

$$\theta + \lambda + \mu = 1$$

Where G_i is score of i_{st} endpoint. Especially, selection is performed independently between high-priority and low-priority groups, if there is no endpoint in expected priority groups, we make the selection among other priority groups, if there are one or two endpoints, we simply choose all of them as candidates without picking more endpoints from other priority groups. After grading candidates, we have a highest-scored endpoint, a middle-scored and a lowest-scored endpoint. For candidate set with one endpoint, the highest-scored, the middle-scored and the lowest-scored endpoint are the same one. For candidate set with two endpoint, the middle-scored and the lowest-scored endpoint are the same one. Then we define priority of request: we regard requests whose requirement level ranks high in at least two factor as high-priority requests, we regard the others as low-priority requests. Based on request priority, we have different values for θ, λ, μ, and different selecting strategy. Table 3 describes the definition of request priority. Table 4 gives a hint to select candidate for different request priority. High-priority requests consume more resource than others, so we choose the highest-scored candidate for them. As for low-priority requests, we choose middle-scored one instead. In this way, we use middle-scored candidates which are healthy enough to handle low-priority requests to reduce workload of high-scored candidates to prevent the situation that high-scored candidate become overhead quickly because requests of all priority rush to them.

3. Rewrite destination of request with the selected endpoint IP and dispatch request.

Type	CPU	Memory	Network	Priority
request1	High	High	High	High
request2	High	High	Low	High
request3	High	Low	High	High
request4	High	Low	Low	Low
request5	Low	High	High	High
request6	Low	High	Low	Low
request7	Low	Low	High	Low
request8	Low	Low	Low	Low

Table 3: request priority

Type	θ	λ	μ	Selected endpoint
request1	0.4	0.3	0.3	highest-scored
request2	0.4	0.4	0.2	highest-scored
request3	0.4	0.2	0.4	highest-scored
request4	0.6	0.2	0.2	middle-scored
request5	0.2	0.4	0.4	highest-scored
request6	0.2	0.6	0.2	middle-scored
request7	0.2	0.2	0.6	middle-scored
request8	0.4	0.3	0.3	middle-scored

Table 4: selecting strategy

It is worth mentioning that the algorithm is based on dynamic data provided by load monitor except for request data. The resource consumption, which is a static and stable property of request, is seemed to remain almost the same for a rather long time, so we just evaluate it for once in load monitor. As for dynamic system load of nodes and endpoints, load monitor keeps on collecting load all the time and reports them to loadbalancer at intervals.Loadbalancer makes load-balancing decision based on the newest verison of data it received from load monitor until the next interval arrives. The length of interval has a great influence on the overhead and performance of loadbalancer. Whenever a new interval time arrives, loadbalancer has to recompute the endpoint load to get real-time load so with the increasement of interval, it costs less overhead to compute but has worse performance because the load are not real-time enough. We set the interval to 5 second to make a trade-off between performance and overhead.

4 Evaluation

This section presents a set of evaluations to show the performance of dynamic solution compared with the original ingress controller. We establish a kubernetes cluster to be the base of the whole evaluation. Then we emulate five workload patterns and run them for twice, one with the original ingress controller equiped, the other with our modified ingress controller equiped to compare their performance. In summary, our results show that the modified ingress controller achieves better performance in most web cases. For independent web applications, the modified ingress controller can balance load well even when there is a great difference between resource consumption of requests. Furthermore, in scaling case, modified ingress controller succeeds to split more load to the new idle endpoints, helping reduce pressure of high-load endpoints and improving RPS. For dependent web applications, modified ingress controller helps protecting endpoints from being influenced by node load, resource race against other endpoints and dependencies to other applications. Modified ingress controller addresses these cases with extra 1/3 overhead. The original ingress controller consumes about 32 35 millicore to synchronize endpoints changes while modified ingress controller consumes 10 more millicore for load monitor and dynamic load balancing.

4.1 Configuration

Our evaluate platform consist of a tsung server, a original ingress controller, our modified ingress controller, several test applications and datebases. All these are deployed as pods on kubernetes. Our kubernetes cluster is made up of a master and five nodes. The hardware configuration of master is Intel Core processor i7 6700 with 4 CPU cores(3.4Ghz), 64GB system memory and a 1TB HDD disk. Nodes share the same configuration: Intel Xeon processor E5 2620 with 8 cores (2.10Ghz), 64GB system memory and a 1TB HDD disk. Many system components are deployed on the master, because it is responsible for managing the whole cluster, making it in high-load state, so we do not put any workload on master. The tsung server is a request generator to emulate workload pattern for comparing performance, and we deploy it alone on node1 without putting any other application on node1 to make sure its perfomance on generating requests is not influenced by other unrelated factors during the evaluation. Similarly, to gurantee the pure performance, we let ingress controller deployed alone on node2. The test application is a java web application connected to a database, it has several apis exposed, requests of different api have diverse consumption on CPU,memory,network and database. The database is a MySQL instance accessed by the test applications. We deploy test applications on node3, node4, node5 and database on node6. With different strategies to generate workload, different number and placement of test application and database, we emulate fix workload pattern to compare performance.

4.2 Workload Pattern

In this section, we evaluate four workload patterns to evaluate performance of modified ingress controller compared with original ingress controller. The workload patterns include initialization pattern, scaling pattern, node-blocking pattern and database-blocking pattern with four different setting of endpoint placement, scaling strategies and workload composition.

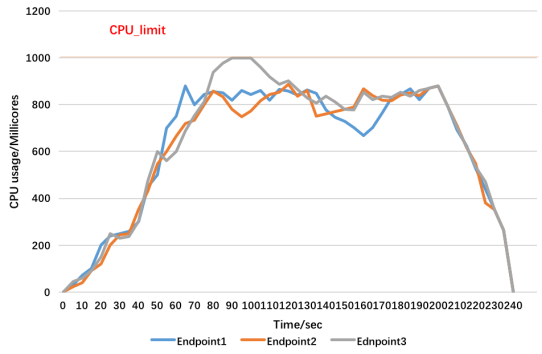

Figure 5: original controller under initialization pattern

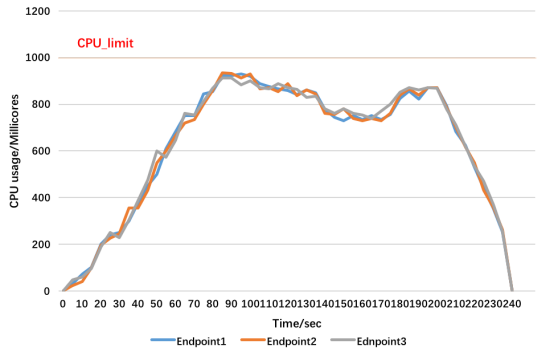

Figure 6: modified controller under initialization pattern

Initialization pattern stands for the initial state when test applications are just deployed: we first deploy three replicas of test application, that is endpoint1 in node3, endpoint2 in node4 and endpoint3 in node5 to prevent that they are impacted by each other. After the deployment, we generate requests of all kinds at a fixed speed by tsung server to access test applications through ingress controller. For convenience, we use the cpu statistics to present performance. Figure 5 presents the performance of original ingress controller under initialization pattern and figure 6 shows the performance of modified ingress controller. Result shows that for the original controller, load of endpoints become unstable and endpoint3 once becomes overhead(its cpu usage excceed its cpu limit) during 80s-100s, causing several requests return 503 to tsung server, it confirms that the default round-robin algorithm only performs well when resource consumption of requests are almost the same, which does not match the real circumstance initialization pattern describes: different kinds of requests have diverse resource requirements, under this circumstance, if we dispatch requests evenly to endpoints, some endpoints can become overhead presumably. Therefore, 503-request appears in evaluation for original ingress controller. Figure 6 shows that with the effort of modified ingress controller, there is no more overhead endpoint, load of endpoints are more even instead. More and more, table 5 points out that original ingress controller finishes 8892 reqeust, with 1108 503-requests, the modified ingress controller finishes 10000 requests with no 503-request, and the total number of requests is 10000.

	Finished	503	Total
Original	8892	1102	10000
Modified	10000	0	10000

Table 5: Request summary of initialization pattern

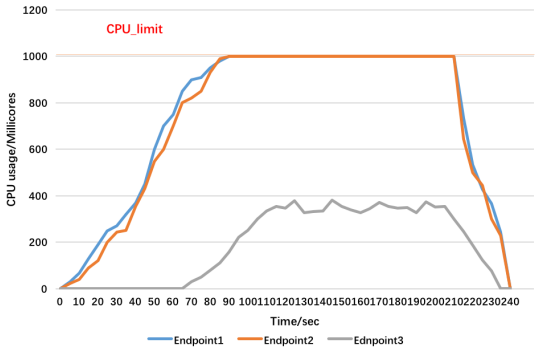

Figure 7: original controller under scaling pattern

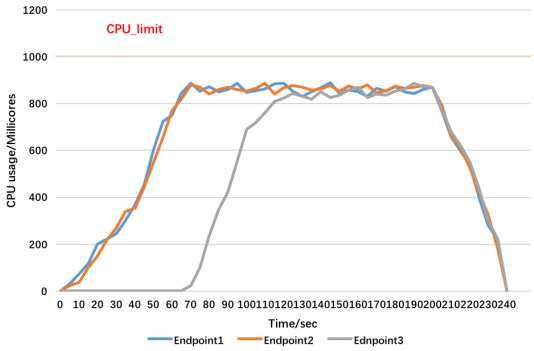

Figure 8: modified controller under scaling pattern

Scaling pattern happens when administrator decides to scale up test applications because the workload is too large for current number of endpoint to handle. To emulate this pattern, we first deploy endpoint1 in node3 and endpoint2 in node3 and generate high workload to keep these two endpoints in high-load state but not overload. Then we deploy endpoint3 in node5 and increase the workload to the level that can keep three endpoints in high-load state ideally. Figure 7 presents the performance of original ingress controller under scaling pattern and figure 8 shows the performance of modified ingress controller. Result shows that that when we increse workload at 65s, original ingress controller failes to dispatch appropriate workload to endpoint3, which should be assigned more workload to help endpoint1 and endpoint2. On the contrary, the modified ingress controller succeeds to lead more workload to endpint3, as a result, no endpoint becomes overhead. Table 6 points out that original ingress finishes xxx requests, with xxx 503-request while the modified ingress controller finishes xxx request with no 503-request, and the total number of reqeust is xxx.

	Finished	503	Total
Original	5844	2956	8800
Modified	8800	0	8800

Table 6: Request summary of scaling pattern

Node-blocking pattern is established to evaluate performance of ingress controllers when their hosts are in high-load state. Firstly, we deploy endpoint1 in node3 and endpoint2 in node4. Then we generate workload for node3 to keep it on high-load state, leaving only enough resource for endpoint1. After that, we generate high workload to keep these two endpoints in high-load state but not overload and by comparing the RPS, we obtain performing disparity between original ingress controller and modified ingress controller. Through the comparison between figure **??** and figure **??**, we find that it costs the original one about xxx second to handler xxx requests while it costs the modified one only YYY seconds, which demonstrates the fact that load of host can affect the performance of endpoints. That is because the default isolation mechanism of docker, which is the container engine of kubernetes, can not support node-blocking pattern well. Kubernetes provides *resource.request* and *resource.limit* for users to allocate resource for their applications, however it does not gurantee the monopolization of resource. For example, the request limit and request of cpu are implemented by a combination of *cgroup* commands: cfs_quota_us,cfs_period_us and *cpushare*. These commands insulates cpu through restricting the length of cpu time slice without binding applications to specific cpus, so when preemption happens, the inevitable context switch brings extra overhead. The modified ingress controller reduce context switch by taking load of host into consideration when balancing load so it gets a higher RPS.

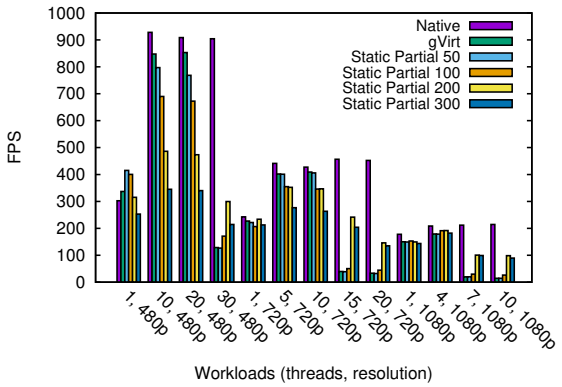

Figure 9: gHyvi with static reconstruction policy

Interrelationship pattern describes situation when several endpoints are deployed in the same node, we set up this pattern to figure out the ability of ingress controller to eliminate interrelationship between these endpoints. To emulate this pattern, we put endpoint1, endpoint2, endpoint3 in node3, endpoint4 in node4 and endpoint5 in node5. Then we generate high workload to keep these endpoints in high-load state. Figure **??** shows that in original case, loads of endpoint1, endpoint2 and endpoint3 are higher than other endpoints at the same time, and finally it provides xxx RPS. That is because endpoints of the same application share the same resource feature, so if endpoints are deployed into the same node, their race for resource becomes fierce, thus they impact on performance of each other. Figure **??** demonstrates that the modified ingress controller succeeds to reduce race and provide xxx RPS.

Database-blocking pattern aims to figure out whether ingress controller can perform well under the micro service background. Under this circumstance, web applications tend to rely on each other rather than finish jobs independently, so QOS is not only decided by application itself, but also other applications it relies on. For convenience, we choose database-accessing applications as study case and set up database-blocking pattern: First we deploy endpoint1 in node3, endpoint2 in node4, endpoint3 in node5, database1 and database2 in node6, then we configure endpoint1, endpoint2 to connect to database1 and endpoint3 to database2. Finally, we generate workload to keep these endpoints in high-load state but not overhead. By comparing RPS, we get their performing disparity. Figure **??** shows that original ingress controller failes to divide load of database1 to datebase2

which comes from endpoint1 and endpoint2. As a result, database1 bears more load than database2 all the time, pulling down the entire RPS. As figure **??** shown, modified ingress controller succeeds to balance load across database1 and database2, providing xxx RPS.

5 Related Work

5.1 GPU Benchmarks

Since GPUs are used for acceleration of general purpose computing, some benchmarks have been implemented for evaluating their performance. Rodinia [5] is a benchmark suite for heterogeneous computing. It aids architects in the study of emerging platforms such as GPUs. Rodinia includes applications and kernels that target multi-core CPU and GPU platforms. And Parboil [19] is a set of throughput computing applications useful for studying the performance of throughput computing architecture and compilers. It collects benchmarks from throughput computing application researchers in many different scientific and commercial fields including image processing, bio-molecular simulation, fluid dynamics, and astronomy.

Unfortunately, the benchmarks above are not available for Intel's GPU now. Meanwhile, GPU's media performance has become a big concern for service providers. However, there is no benchmark specifically for this kind of workload. So, this paper proposes GMedia, a media transcoding benchmark based on Intel's MSDK.

5.2 GPU Virtualization

Though virtualization has been studied extensively in recent years, GPU virtualization is still a nascent area of research. Typically, there are four ways to use GPU in a Virtual Machine (VM): I/O pass-through, device emulation, API remoting, and mediated pass-through.

A naive way to use GPU in virtualized environment would be to directly pass through the device to a specific VM [13, 7]. However, the GPU resources are dedicated and cannot be multiplexed.

Device emulation, similar to binary translation in CPU virtualization, is impractical. GPUs, unlike CPUs, whose specifications are not well documented, vary between vendors [8]. Emulating GPUs from different vendors requires vast engineering work. Notably, following up the new GPU hardware would make it a nightmare to maintain the codebase.

API remoting is widely used in commercial softwares such as VMWare and VirtualBox, and has been studied throughout many years. By using API remoting, graphic commands are forwarded from guest OS to host. VMGL [16] and Oracle VirtualBox [2], both based on Chromium [14], replace the standard OpenGL library in Linux Guests with its own implementation to pass the OpenGL commands to VMM. Nonetheless, forwarding OpenGL commands is not considered a general solution, since Microsoft Windows mainly uses their own DirectX API. Whether forwarding OpenGL or DirectX commands, it would be difficult to emulate the other API. gVirtuS [10], VGRIS [17], GViM [12], rCUDA [9] and vCUDA [18] use the same manner to forward CUDA and OpenCL commands, solving the problem of virtualizing GPGPU applications.

VMware's products consist of a virtual PCI device, SVGA II card [8], and the corresponding driver for different operating systems. The emulated device acts like a real video card which has registers, graphics memory and a FIFO command queue. All accesses to the virtual PCI device inside a VM is handled on the host side, by a user-level process, where the actual work is performed. Moreover, they have designed another graphic API called SVGA3D. The SVGA3D protocol is similar to Direct3D and shares a common abstraction. The purpose of SVGA3D is to eliminate the commands for a specific GPU. Meanwhile, a GPU can also emulate the missing features by SVGA3D protocol, which provides a practical portability for their products.

Recently, two full GPU virtualization solutions have been proposed, i.e., gVirt of Intel [21] and GPUvm [20], respectively. gVirt is the first open source product level full GPU virtualization solution in Intel platforms. gVirt presents a vGPU instance to each VM which allows the native graphics driver to be run in VM. The shadow page table is updated with a coarse-grained model, which could lead to a performance pitfall under some video memory intensive workloads, such as media transcoding.

GPUvm presents a GPU virtualization solution on a NVIDIA card. Both para- and full-virtualization were implemented. However, full-virtualization exhibits a considerable overhead for MMIO handling. The performance of optimized para-virtualization is two to three times slower than native. Since NVIDIA has individual graphics memory on the PCI card, while the Intel GPU uses part of main memory as its graphics memory, the way of handling memory virtualization is different. GPUvm cannot handle page faults caused by NVIDIA GPUs [11]. As a result, they must scan the entire page table when translation lookaside buffer (TLB) flushes. As gHyvi allocates graphics memory within the main memory, VMM can write-protect the page tables to track the page table modifications. This fine-grained page table update mechanism mitigates the overhead incurred by the Massive Update Issue.

NVIDIA GRID [1] is a proprietary virtualization solution from NVIDIA for Kepler architecture. However, there are no technical details about their products avail-

able to the public.

5.3 Memory Virtualization

One important aspect in GPU virtualization is memory virtualization, which has been thoroughly researched. The software method employs a shadow page table to reduce the overhead of translating a VM's virtual memory address. This approach could incur severe overhead under some circumstances. Agesen *et al.* [4] listed three situations where the shadow page table cannot handle well: the hidden page fault, address space switching, and the tracing page table entries. They also pointed out some optimization techniques, such as the trace mechanism and eager validating. Unfortunately, it is hard to trade off these mutually exclusive techniques. Therefore, AMD and Intel have added the hardware support for memory virtualization. All three overheads previously listed before can be eliminated, but it is not the silver bullet, a TLB miss punishment is higher in the hardware solution. In the classical VMM implementations, VMM employs a trace technique to prevent its shadow PTEs from becoming inconsistent with guest PTEs, i.e. updating shadow page table strictly after the guest page table is modified. Typically, VM trace uses write-protection mechanism, which can be the source of overhead. This technique is similar to the current gVirt's strict page table shadowing mechanism, which frequently traps and emulates the page faults of the shadow page table, and it causes overhead. gHyvi removes the write-protection from shadow page table to eliminate the overhead caused by excessive trap-and-emulation, taking advantage of the GPU programming model [3].

6 Conclusion and Future Work

gHyvi is an optimized full GPU virtualization solution, based on the Xen hypervisor, with the adaptive hybrid page table shadowing scheme, which improves performance for workloads with the Massive Update Issue when compared to gVirt. To address this issue, this paper provides a hybrid page table shadowing scheme, i.e., strict and relaxed page table shadowing, to provide an optimized full GPU virtualization based on Xen hypervisor for Intel GPUs. gHyvi combines these two page table shadowing mechanisms to reduce VM-exits to the hypervisor. Further, gHyvi automatically switches page table between them by detecting GPU's current workloads, potentially showing significantly improvement to gVirt's performance for workloads with the Massive Update Issue. In order to decide what type of the page need to be reconstructed, four reconstruction policies are introduced. By running the same testcase through the four policies, the dynamic segmented partial reconstruction policy performs the best.

For future work, we will adapt gHyvi to support KVM [15] when gVirt for KVM is ready. Additionally, gHyvi will be released in the open source community soon. We will focus on the areas of portability, scalability, and scheduling issues. With previous GPU command scheduling methods, such as VGRIS and Pegasus [6], we will investigate the low level access pattern of massive page table modification with the detailed analysis of the performance bottleneck of high level applications. We hope this optimized full GPU virtualization solution gives insight into designing the support of efficient distributed systems for GPU acceleration applications.

7 Acknowledgements

We thank our shepherd Dan Tsafrir, Haibo Chen, and the anonymous reviewers for their insightful comments. This work was supported by National Science and Technology Major Project (No. 2013ZX03002004), National R&D Infrastructure and Facility Development Program (No. 2013FY111900), NRF Singapore CREATE Program E2S2, and Shanghai Key Laboratory of Scalable Computing and Systems. Prof. Haibing Guan is the corresponding author.

References

[1] Nvidia grid: Graphics-accelerated virtualization. `http://www.nvidia.com/object/grid-technology.html`.

[2] Oracle vm virtualbox. `https://www.virtualbox.org/`.

[3] ADAMS, K., AND AGESEN, O. A comparison of software and hardware techniques for x86 virtualization. *ACM Sigplan Notices 41*, 11 (2006), 2–13.

[4] AGESEN, O., GARTHWAITE, A., SHELDON, J., AND SUBRAHMANYAM, P. The evolution of an x86 virtual machine monitor. *ACM SIGOPS Operating Systems Review 44*, 4 (2010), 3–18.

[5] CHE, S., BOYER, M., MENG, J., TARJAN, D., SHEAFFER, J. W., LEE, S.-H., AND SKADRON, K. Rodinia: A benchmark suite for heterogeneous computing. In *Workload Characterization, 2009. IISWC 2009. IEEE International Symposium on* (2009), IEEE, pp. 44–54.

[6] DEELMAN, E., SINGH, G., SU, M.-H., BLYTHE, J., GIL, Y., KESSELMAN, C., MEHTA, G., VAHI, K., BERRIMAN, G. B., GOOD, J., ET AL. Pegasus: A framework for mapping complex scientific workflows onto distributed systems. *Scientific Programming 13*, 3 (2005), 219–237.

[7] DONG, Y., DAI, J., HUANG, Z., GUAN, H., TIAN, K., AND JIANG, Y. Towards high-quality i/o virtualization. In *Proceedings of SYSTOR 2009: The Israeli Experimental Systems Conference* (2009), ACM, p. 12.

[8] DOWTY, M., AND SUGERMAN, J. Gpu virtualization on vmware's hosted i/o architecture. *ACM SIGOPS Operating Systems Review 43*, 3 (2009), 73–82.

[9] DUATO, J., PENA, A. J., SILLA, F., MAYO, R., AND QUINTANA-ORTÍ, E. S. rcuda: Reducing the number of gpu-based accelerators in high performance clusters. In *High Performance Computing and Simulation (HPCS), 2010 International Conference on* (2010), IEEE, pp. 224–231.

[10] GIUNTA, G., MONTELLA, R., AGRILLO, G., AND COVIELLO, G. A gpgpu transparent virtualization component for high performance computing clouds. In *Euro-Par 2010-Parallel Processing*. Springer, 2010, pp. 379–391.

[11] GOTTSCHLAG, M., HILLENBRAND, M., KEHNE, J., STOESS, J., AND BELLOSA, F. Logv: Low-overhead gpgpu virtualization. In *High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing (HPCC_EUC), 2013 IEEE 10th International Conference on* (2013), IEEE, pp. 1721–1726.

[12] GUPTA, V., GAVRILOVSKA, A., SCHWAN, K., KHARCHE, H., TOLIA, N., TALWAR, V., AND RANGANATHAN, P. Gvim: Gpu-accelerated virtual machines. In *Proceedings of the 3rd ACM Workshop on System-level Virtualization for High Performance Computing* (2009), ACM, pp. 17–24.

[13] HIREMANE, R. Intel virtualization technology for directed i/o (intel vt-d). *Technology@ Intel Magazine 4*, 10 (2007).

[14] HUMPHREYS, G., HOUSTON, M., NG, R., FRANK, R., AHERN, S., KIRCHNER, P. D., AND KLOSOWSKI, J. T. Chromium: a stream-processing framework for interactive rendering on clusters. In *ACM Transactions on Graphics (TOG)* (2002), vol. 21, ACM, pp. 693–702.

[15] KIVITY, A., KAMAY, Y., LAOR, D., LUBLIN, U., AND LIGUORI, A. kvm: the linux virtual machine monitor. In *Proceedings of the Linux Symposium* (2007), vol. 1, pp. 225–230.

[16] LAGAR-CAVILLA, H. A., TOLIA, N., SATYANARAYANAN, M., AND DE LARA, E. Vmm-independent graphics acceleration. In *Proceedings of the 3rd international conference on Virtual execution environments* (2007), ACM, pp. 33–43.

[17] QI, Z., YAO, J., ZHANG, C., YU, M., YANG, Z., AND GUAN, H. Vgris: Virtualized gpu resource isolation and scheduling in cloud gaming. *ACM Transactions on Architecture and Code Optimization (TACO) 11*, 2 (2014), 17.

[18] SHI, L., CHEN, H., SUN, J., AND LI, K. vcuda: Gpu-accelerated high-performance computing in virtual machines. *Computers, IEEE Transactions on 61*, 6 (2012), 804–816.

[19] STRATTON, J. A., RODRIGUES, C., SUNG, I.-J., OBEID, N., CHANG, L.-W., ANSSARI, N., LIU, G. D., AND HWU, W.-M. Parboil: A revised benchmark suite for scientific and commercial throughput computing. *Center for Reliable and High-Performance Computing* (2012).

[20] SUZUKI, Y., KATO, S., YAMADA, H., AND KONO, K. Gpuvm: why not virtualizing gpus at the hypervisor? In *Proceedings of the 2014 USENIX conference on USENIX Annual Technical Conference* (2014), USENIX Association, pp. 109–120.

[21] TIAN, K., DONG, Y., AND COWPERTHWAITE, D. A full gpu virtualization solution with mediated pass-through. In *Proc. USENIX ATC* (2014).