DOI 10.1007/s11036-015-0664-6

Cloud-Integrated Cyber-Physical Systems for Complex Industrial Applications

Zhaogang Shu [1] · Jiafu Wan [2] · Daqiang Zhang [3] · Di Li [2]

Published online: 28 November 2015

Abstract Recently, with many advances in wireless sensor networks, big data, mobile and cloud computing, Cyber-Physical Systems (CPS) can tightly couple cyber space with the physical world better than ever before. Also, the cloud-based systems can provide massive storage resources and low-cost computing as well as the flexibility of customizing the operating environment to Complex Industrial Applications (CIA). In our view, Cloud-integrated CPS (CCPS) will open the door to allow previously unachievable application scenarios to be built, deployed, managed and controlled effectively. In this paper, we propose a novel architecture of CCPS (termed CCPSA) and outline the enabling technologies for CIA. Then, we dissect three potential challenges and provide solutions from the perspective of CIA, including virtualized resource management techniques, the scheduling of cloud resources, and life cycle management. We hope this paper can provide insight and a roadmap for future research efforts in the emerging field of CCPS.

✉ Jiafu Wan
mejwan@scut.edu.cn

Zhaogang Shu
zhaogang.shu@gmail.com

Daqiang Zhang
dqzhang@ieee.org

Di Li
itdili@scut.edu.cn

[1] Fujian Agriculture and Forestry University, Fuzhou, China

[2] South China University of Technology, Guangzhou, China

[3] Tongji University, Shanghai, China

Keywords Cyber-physical systems · Cloud computing · Industrial wireless networks · Complex industrial applications · Big data

1 Introduction

Cyber-Physical Systems (CPS) are reliable and evolvable networked time-sensitive computational systems integrated with physical processes, and are being widely used in many critical areas, such as manufacturing, traffic control and safety [1, 2]. Recently, CPS has made great strides in some aspects (e.g., safety and security [3, 4], abstraction and verification [5], modeling [6], data processing [7], and control [8]), but it requires more attention on other approaches. For example, it is necessary that the design of an innovative methodology for closely combining CPS with cloud computing meets the requirements of industry 4.0, i.e., the Complex Industrial Applications (CIA).

Since cloud computing has the ability to provide a flexible stack of massive computing, storage and software services in a scalable and low-cost manner, it has been widely used [9, 10]. Nevertheless, it is generally known that the majority of current cloud systems and the corresponding techniques primarily focus their attention on Internet-based applications. The CIA brings in challenges to cloud computing, since they are significantly distinguishable from those service-oriented Internet-based applications due to their inherent characteristics, such as workload variations, real-time scheduling, and adaptive resource management.

With advances in embedded design, wireless sensor networks, mobile computing and big data, it is an inevitable trend in designing large-scale complicated systems by integrating CPS with cloud computing, and CIA is not an exception. Cloud-integrated CPS (CCPS) refers to virtually representing physical system components such as sensors, actuators, robots and other devices in clouds. It involves accessing (e.g., monitoring, actuating and navigating) those physical components through their virtual representations, and processing the large amount of data collected from physical components in clouds in a scalable, real-time, efficient, and reliable manner. Particularly, integrating cloud computing techniques (e.g., virtualization) with CPS techniques (e.g., real-time scheduling, and adaptive resource management and control) will allow previously unachievable systems such as cloud-integrated manufacturing and cloud-integrated vehicles to be deployed effectively.

In this paper, we explore the simplified architecture and the enabling technologies in the design of CCPS for CIA, review several challenges and provide potential solutions to improve the Quality of Service (QoS) of the related CCPS. We highlight the insights and contributions as follows.

The simplified architecture of cloud and CPS: By incorporating the dynamic interactions between the cloud and CPS, we propose a novel CCPS Architecture (termed CCPSA) to provide flexible services and applications for CIA.

The challenges and solutions of CCPS for CIA: We carefully choose the challenges from the perspective of CIA and give possible solutions, including virtualized resource management techniques, scheduling of cloud resources, and Life Cycle Management (LCM).

The remainder of the paper is organized as follows. Section II reviews the related work on CPS and cloud computing. In Section III depicts CCPSA and its enabling technologies in detail. In Section IV, we discuss key technology challenges for CCSPA and corresponding solutions, which would be proven effective through practice and experiments. Finally, Section V concludes this paper and provides an outlook.

2 Recent Advances in CPS and Cloud

In this section, we briefly review the recent advances related to CPS and cloud computing, and analysis results are summarized.

As mentioned above, CPS has made significant achievements recently. In Ref. [11], R. Rajkumar proposed that the technical challenges of CPS would be overcome in the near future, which would result in a science of CPS and new technological solutions. In this way, practical, affordable, and reliable CPS would be deployed in many domains. In Ref. [5], Rajhans et al. studied the abstraction and verification of the procedure, and an architectural framework for CPS was proposed by using structural and semantic mappings to ensure consistency and enable system-level verification. In Ref. [12], Vamvoudakis et al. designed a game-theoretic approach to estimate a binary random variable based on sensor measurements that may have been corrupted by a cyber attacker. For the control of CPS, especially wireless control and event-based control, Demirel et al. [13] studied wireless control loops with sensor measurements communicated over an unreliable and energy-constrained multi-hop wireless network. As a result, a modular design method was developed that jointly optimized packet forwarding policies and control commands. In Ref. [14], an event-based state estimation problem was studied where the sensors triggered their transmission to a fusion node only if their associated measurement prediction variance exceeded a certain threshold. In addition, many application scenarios of CPS can be found in Ref. [15–17].

Cloud computing has been widely used in recent years. So far, it has achieved abundant research results. Baliga et al. [18] provided a comprehensive energy consumption analysis of cloud computing. The analysis considered both public and private clouds and included energy consumption in switching and transmission as well as data processing and data storage. Warneke et al. [19] exploited dynamic resource allocation for efficient parallel data processing in the cloud. Khazaei et al. [20] analyzed the pool management scheme for cloud computing centers. In Ref. [21], network virtualization and software-defined networking for cloud computing were summarized. Xiao et al. [22] adopted an attribute-driven methodology to systematically study the security and privacy issues in cloud computing. In addition to these results, the most popular cloud computing platforms, such as Amazon, Google, and IBM, have appeared in various domains, such as intelligent transportation systems [9], and the manufacturing industry [15].

Most of the literature on cloud and CPS, except for Refs. [2, 9], have made contributions from the perspective of separation. In Ref. [2], a context-aware vehicular CPS with cloud support was proposed, and two crucial service components (vehicular social networks and context-aware vehicular security) were introduced. In Ref. [9], Wan et al. designed a new architecture (termed VCMIA) by integrating vehicular CPS with mobile cloud computing, which can provide mobile services for potential users such as drivers and passengers to access mobile traffic clouds.

From the above, we can see that advances in wireless communication techniques, mobile and embedded computing, and distributed network control are boosting a growing interest in the design, development, and deployment for the emerging applications of CIA. This leads to an increasing evolutionary tendency that traditional industrial applications will migrate to CCPS for CIA.

3 CCPS for CIA: an inevitable trend

The work in Ref. [23] pointed out that a CPS for CIA requires three levels (see Fig. 1): 1) the physical objects; 2) data models of the mentioned physical objects in a network infrastructure; and 3) services based on the available data. Depending on the cloud platform, the components and other entities in industrial production would be assigned their own identities in the network. They could either negotiate with each other or could be interconnected and simulated. In this section, we will propose a CCPS Architecture (termed CCPSA) for CIA and discuss its enabling technologies.

3.1 Proposed architecture: CCPSA

CCPS can be divided into three different domains: network-centric CPS, cloud-centric CPS, and data-centric CPS, corresponding to communications, control, and computation requirements of industrial development and deployment. Figure 2 shows the proposed CCPSA for CIA.

From the perspective of LCM, through social computing, the new industrial mode can transform traditional enterprises into intelligent enterprises that can actively sense and respond to customers' individualized needs on a massive scale. The key to the success of this new mode is to effectively integrate social demands and production capacity in real time. Therefore, it is imperative to combine the two emerging fields of social computing and industrial production, as well as to seamlessly connect the Internet to logistic networks with industrial robots and 3D printing based manufacturing networks, so that customers can also participate fully in the whole life cycle of the production processes. In Fig. 1, the dotted line shows the process from social requirement to the final product.

From the perspective of the architecture, intelligent factories can autonomously implement the productive process with the support of new devices and technologies, such as industrial robots, Industrial Wireless Networks (IWN), and advanced coordination mechanisms. For example, in order to reduce the energy consumed by a vehicle body assembly line, the idle robots can be powered down during breaks in production by sensing each other's statuses. Such a flexible assembly line could meet individual customer requirements. In the manufacturing process, all the data (e.g., device status, product information, and raw material information) can be forwarded to an industrial cloud for carrying out data analysis and then providing a personalized service.

As mentioned, since CIA has inherent features (e.g., workload variations, real-time scheduling, and adaptive resource management), current cloud systems cannot often meet the application requirements of CIA. To address these problems, it is necessary to develop efficient virtualization techniques to improve resource management, feature-aware multi-dimensional resource scheduling, etc. By achieving major breakthroughs in models, methods and techniques, we could establish the resource management and scheduling systems targeting CIA in clouds. In the application layer, some innovative

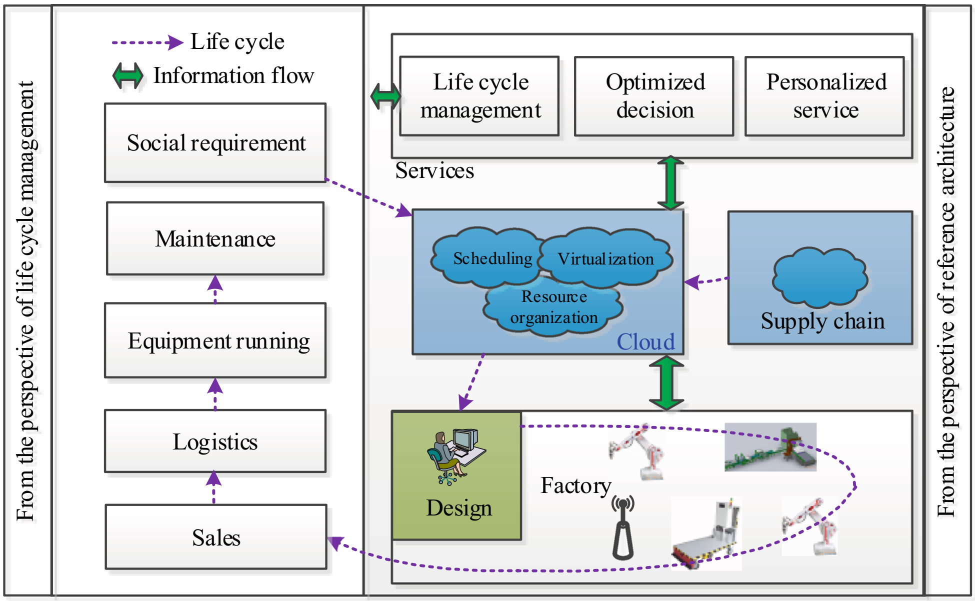


Fig. 1 The Proposed CCPSA for CIA

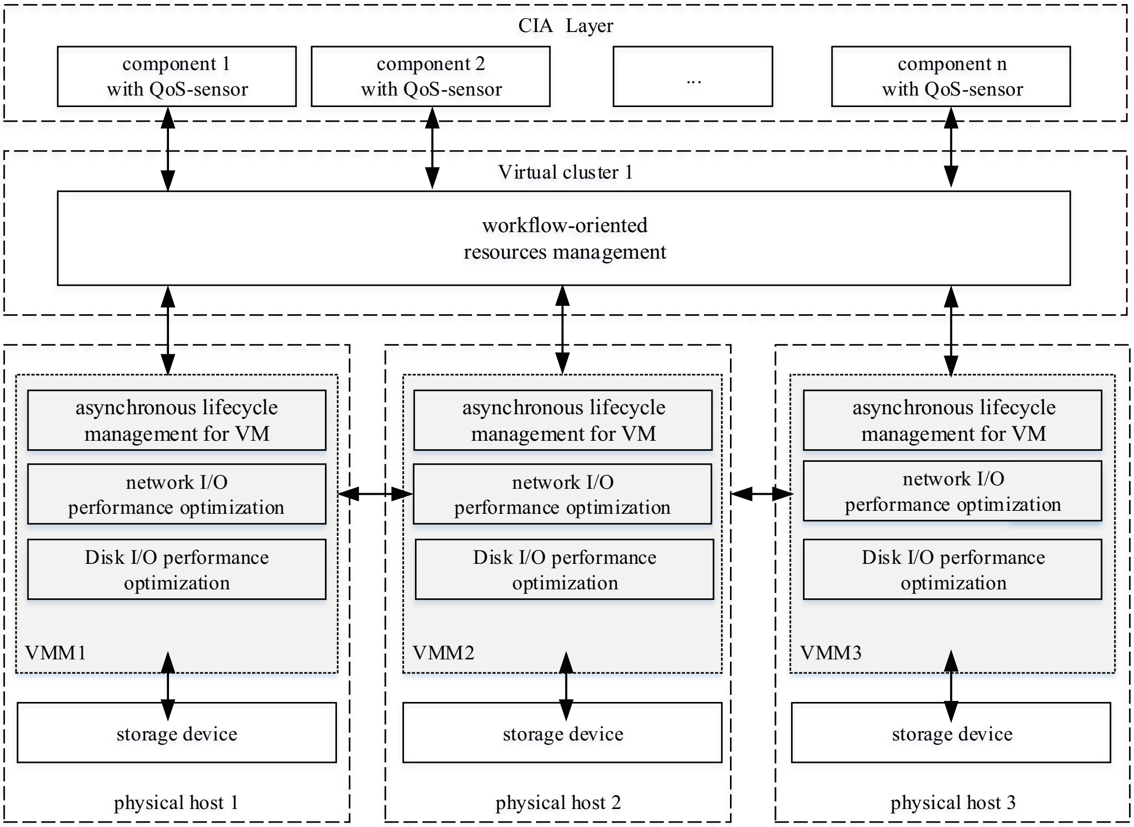

CIA Layer
component 1 with QoS-sensor
component 2 with QoS-sensor
...
component n with QoS-sensor
Virtual cluster 1
workflow-oriented resources management
asynchronous lifecycle management for VM
network I/O performance optimization
Disk I/O performance optimization
VMM1
VMM2
VMM3
storage device
physical host 1
physical host 2
physical host 3

Fig. 2 Virtualized resource management framework for CIA in cloud environments

services (e.g., LCM) and decisions (e.g., production guidance) are also provided by designing the big-data-based applications.

3.2 The enabling technologies for CIA

As pointed out in Ref. [23], the revolution of CIA is not necessarily the technical realization but the new horizon of business models, services, and individualized products, so the enabling technologies of CIA are very important to strengthen the functionality and performance of the whole CIA system. The novel designs of the CIA system are challenging, which requires the significant optimization, control, or even reconstruction of all layers and components in the system, as stated in Ref. [24, 25]. Therefore, various research studies have focused on specific aspects of CIA, which can be classified by seven components, Embedded Control Systems (ECS), IWN, AGV & 3D printing, industrial cloud, industrial big data, social computing, and system integration and optimization decision. These features are illustrated in Table 1 (since we perform the research in this paper from the perspective of information technology, the AGV & 3D printing are not included in the table).

Although many achievements have been made on CIA, there are still many new issues and challenges in the context of CCPS we will focus three aspects (virtualized resource management techniques, scheduling of cloud resources, and LCM) to state the challenges and possible solutions in the next section.

4 Challenges and possible solutions for CIA

4.1 Virtualized resource management techniques for CIA

Cloud computing centers usually provide thousands of virtual computing nodes and virtual storage devices via virtualization technology, its resource organization mode can be regarded as a group of virtual nodes and virtual storage devices that compose a big resource pool, and resource scheduling algorithms will allocate these resources to the corresponding tasks according to various application requirements. Compared to traditional High Performance Computing (HPC), cloud computing has the unique advantage of satisfying the requirements for CIA in cost reductions, high maintainability and so on, but it is still difficult to achieve ideal performance because of the particularity of the CIA.

Firstly, CIA often requires large amounts of calculation and low latency communication, but existing cloud computing resource organization modes are based on

Table 1 A summary of the enabling techniques of all components of the CIA system

Components	Topics	Future research targeting on CCPS for CIA
Embedded control systems	Model-based design: [26, 27]	Interactions and co-design of cyber- and physical sub-systems
	Resource management: [28]	Collaborative energy-saving design of hardware and software
	Integrated control & scheduling: [29]	Efficient integration among control, communication, and computing
Industrial wireless networks	Standards: [30]	Emerging standards and protocols (e.g. IEEE 802.15.4, 6LoWPAN)
	MAC protocols: [31]	Energy efficiency of IWN
	Routing protocols: [32]	Optimization and control of IWN
Industrial clouds	Scheduling & management: [33]	Scheduling of the cloud resources for industrial applications
	Virtualization: [34]	Virtualized resource management techniques for industrial applications
	Security: [35]	Data security (e.g., transmission, processing and storage)
Industrial big data	Methodologies: [36]	Data-driven control, data visualization, and real-time data retrieval methods
	Platforms & applications: [37]	Case studies, such as prototype platform
Social computing	Theories: [38]	Data clustering and classification targeting on CIA
	Typical platforms: [39]	Case studies, such as prototype platform
System integration & optimization decision	Coordination mechanisms: [40]	Formal methods for interaction behaviors among devices
	Distributed network control: [41]	Coordination control of multi-agent systems with various constraints

virtualization technology, which increase the communication latency greatly, so it conflicts with the performance requirements of CIA. Secondly, CIA contains large amount of components whose resource features are very different, such as resource type and resource capacity. For example, some components may be computationally intensive, which may require more computing resources but less storage resources. On the contrary, other components may be I/O-intensive, which may require more I/O resources but less computing resources. Generally, existing cloud environments contain only approximate isomorphic computing nodes, storage nodes or I/O nodes, which may not be flexible enough to adapt to the above situation. Therefore, the existing virtualized resource management techniques are not sufficient to effectively support CIA to obtain high performance in CCPS.

To satisfy the new requirements (e.g., heterogeneous computing, large amounts of synergy nodes, and low communication latency between virtual nodes) for CIA in cloud environment, we propose a virtualized resource management framework, as illustrated in Fig. 2. This framework adopts a set of vectors to describe computing performance, memory capacity, communication bandwidth and I/O performance, instead of the traditional parameters, such as CPU frequency to mark computing power. With the support of the vectors, the framework can conduct a comprehensive evaluation for all virtualized resources. The virtualized resource management technologies targeting CIA include four aspects: workflow- oriented virtual cluster management technologies, disk I/O optimization techniques in data-intensive environments, network I/O optimization techniques in communication-intensive environments, and virtual machine asynchronous lifecycle management mechanisms. In the following, we will analyze how to carry out the virtualized resource management for CIA.

1. *Workflow-oriented virtual cluster resource management*

The resource requirements of the components of CIA may be analyzed in the workflow diagram according to the characteristics of each component, fully considering performance loss across the physical host, and setting the scale of required resources through heuristic optimization algorithms. When the resources scales are small, the optimal configuration can be obtained through genetic, simulated annealing, or ant colony algorithms. For larger resources scales, the affinity and dependence of all components must be analyzed in the workflow diagram according to the resource feature vectors. For example, in order to obtain the mapping from the virtual CPU to the physical CPU, fuzzy mathematics may be used to get the fuzzy definition of computing capacity requirements. We then take the efficient heuristic algorithm, and finally produce the reasonable computing resources scale. Based on the scale, the collaborative scheduling mechanism can be designed for the virtual CPU across physical hosts. The general steps are as follows. First, we analyze the source of communication latency in the virtual cluster. Second, we prove the correctness of collaborative scheduling in theory. Finally, we implement the collaborative scheduling algorithm.

A workflow is commonly represented by a directed acyclic graph (DAG) $G = (T,D)$ with n nodes (or tasks), where $t_i \in$

T $(1 \leq i \leq n)$ is a workflow task with computation cost $c_i \in R^+$and $d_{i\ ,j\in}D(i \neq j)$ is a dependency between and $t_i$and t_j with an associated communication cost $c_{i \leftrightarrow j} \in R^+$. Commonly, the nodes in the directed acyclic graphs are labeled with their computational cost (e.g. the number of instructions), while edges are labeled with the communication cost between two nodes. A cloud is a set of heterogeneous resources $R = \{r_1, r_2, \ldots, r_n\}$, with associated processing capacities $p_i \in R^+$, connected by network links. Each resource $r_i \in R$ has a set of links $L_i = \{l_{i1}, l_{i2}, \ldots, l_{im}\}(1 \leq m \leq n)$,where $l_{ij} \in R^+$ is the available bandwidth in the link between resource r_1 and r_2. Task scheduling is a function schedule $_{\mathrm{DAG}}$:$T \rightarrow R$, where each task $t_i \in T$ is mapped to a resource $r_i \in R$.

It is worth mentioning that the QoS becomes an important vector for collaborative scheduling algorithms in virtual clusters. Therefore, most components of CIA must be given a feature (U, V) that is quantifying QoS (termed QoS-sensor), which can be used to determine the priority of each component in the scheduling mechanism. Depending on the workflow-oriented virtual cluster management technology, we can effectively improve the performance of CIA in cloud environments.

2. *Asynchronous lifecycle management of virtual machine resources*

The virtual machine lifecycle management includes the creation, startup, configuration, optimization, preservation, restoration, cloning, closure and other activities of virtual machines. Combined with virtual machine template technology, virtual machine lifecycle management can cooperate with the efficient shared storage or store a snapshot in the cloud environment, which can effectively create virtual machines quickly. Through the technology of virtual machine cloning and virtual machine memory recovery, the problem of launching a large number of virtual machines in a short time can be solved. The problem for destruction of virtual machines is relatively simple, but it is necessary to guarantee that virtual machine persistent storage must be protected.

3. *Disk I/O resource optimization in data-intensive environments*

Usually, many intermediate files are produced in the data-intensive CIA workflow, and these intermediate files may be stored on disk through virtual machine I/O interfaces. In order to improve the I/O performance, it is necessary to analyze the characteristics of virtual machine disk access and optimize disk I/O scheduling algorithms, but the premise is not to change the transparency and encapsulation characteristics of virtual machines.

To satisfy I/O requirements, such as response time and deadline, the traditional disk I/O scheduler, such as EDF (Early Deadline First), ignores the locality issues of disk operations during system execution. However, I/O requests in CIA have strong locality relations in data-intensive environments because a cloud platform allocates storage space for different applications from same virtual machines in the same sector or nearby sectors. To resolve this issue, we propose a disk I/O scheduling framework for CIA, called DISF-CIA. The framework is composed of a QoS controller in the backend driver and a locality-aware scheduling algorithm in the native driver. The QoS controller assigns an I/O threshold for each virtual machine to control the throughput of each virtual machine according to its workload and service type. The locality-aware scheduling algorithm is developed based on the deadline modification SCAN algorithm, which is an improved EDF algorithm. Figure 3 shows the architecture of DISF-CIA.

We integrated the DISF-CIA scheme on a Xen-based hypervisor. In order to measure the performance of DISF-CIA, we generated random access patterns and sequential access patterns, and these patterns were run on a Xen-based hypervisor integrated with DISF-CIA Scheme. The hardware environment includes two physical machines and corresponding virtual machines. In our experiments, we generated two data sets that had some characteristics of CIA. The first data set was generated to simulate conditions where the virtual machine runs a sequential access application, such as multimedia streaming servers. The second data set contains many random access I/O requests, which can simulate conditions where CIA accesses many small files. Table 2 shows the details of the generated data sets.

We measured the total number of I/O operations per second (IOPS) between DISF-CIA and traditional EDF schedulers. IOPS is a common performance measurement used to benchmark computer storage devices, such as hard disk drives (HDD). We also compare the number of missed deadline jobs

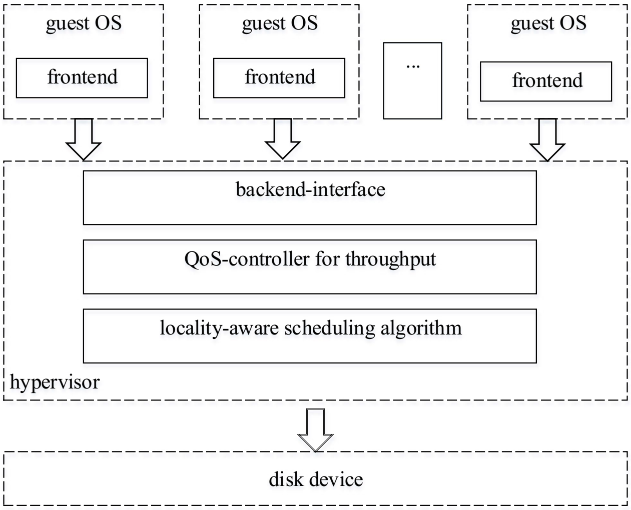


Fig. 3 The Architecture of DISF-CIA

Table 2 The characteristics of data sets

Items	First data set	Second data set
Data access type	Sequential	Random
Total data size	4GB	2GB
Number of data files	10	256
Number of I/O request	327,682	327,682
Number of threads for application task	8	16

and latency. Figures 4 and 5 shows the IOPS comparison for sequential access pattern and random access pattern respectively. Compared with EDF, DISF-CIA improves the IOPS by 5 %-7 % in the sequential access case. For the random access case, the DISF-CIA improves the IOPS by 5 %-8 %. In a comparison of the number of missed deadlines, Fig. 6 shows that the DISF-CIA also obtains a significant I/O performance improvement.

4. *Network I/O resource optimization in communication- intensive environments*

In the CCPS environment for CIA, network I/O communications between virtual nodes happen frequently. There is an alternative to establish performance optimization strategies for parallel application network I/O communication. The key to improve the communication performance of parallel applications is to avoid the unnecessary message polling. To this end, the following three approaches may be considered: a) using blocking polling mechanisms instead of busy polling mechanism; b) improving scheduling algorithms of VMM and publishing scheduling information of the guest OS to the VMM; and c) publishing the scheduling information of VMM to the guest OS. For the first approach, regarding the communication scenario between different virtual machines, the blocking operation should be performed on a file descriptor set, instead of non-blocking operations implemented on the MPI library. This method can avoid the waste of CPU resources on the unnecessary message polling. However, this solution will cause frequent and expensive VCPU context switch, so it may reduce the performance and make the implementation of block polling operations in the shared memory more difficult. In order to solve this problem, another method is proposed to improve performance of communication for the parallel I/O application. When the process that is receiving a message is blocked and ready to cede ownership of the processor, it will inform the VMM to reschedule at an appropriate time.

In order to verify the feasibility of above method, we developed a Resource Scheduling Module (RSM) on Xen, which will publish scheduling information of the guest OS to the VMM and improve the scheduling efficiency of VMM. With the help of RSM, we can analyze the web server performance in a virtualized environment. The following metrics can be used in the measurement study.

a) *Server throughput* (i.e. *the number of requests per second*). One way to measure web server performance is to measure the server throughput in each VM at different workload rates, namely the maximum number of successful requests served per second when retrieving web documents.
b) *Aggregated throughput* (i.e. *the number of requests per second*). We use aggregated throughput as a metric to compare and measure the impact of using different

Fig. 4 The comparison of IOPS for the sequential access pattern

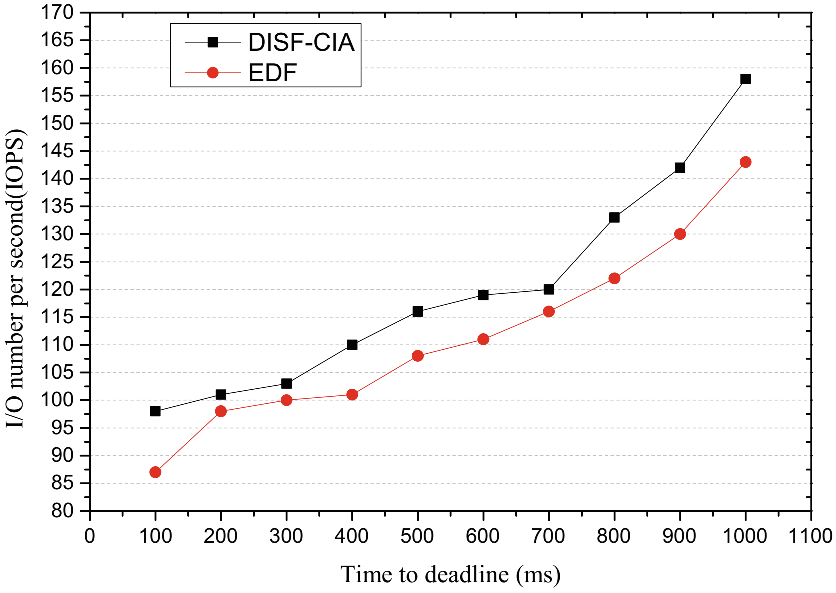
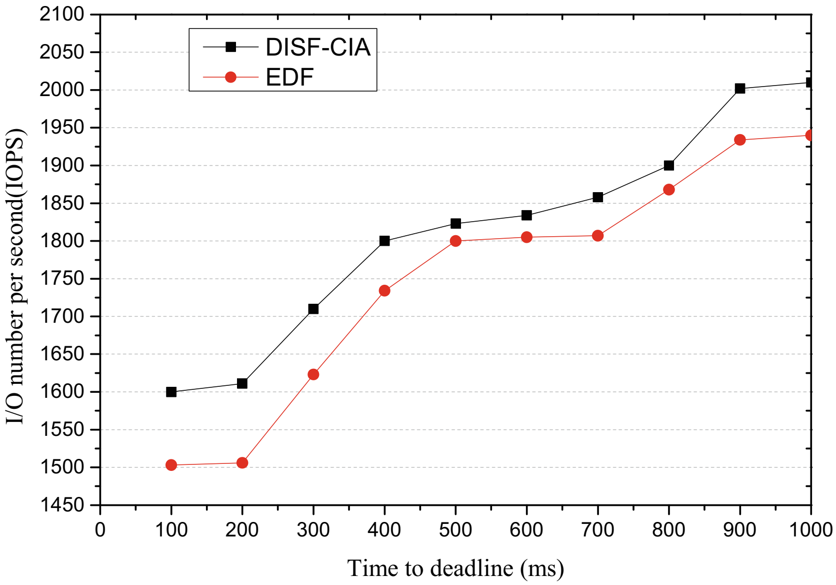


Fig. 5 The comparison of IOPS for the random access pattern



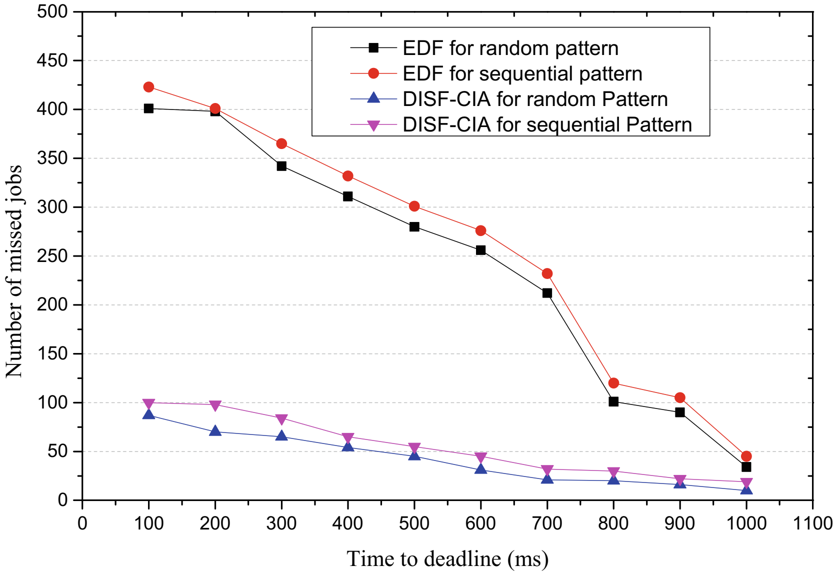
numbers of VMs on the aggregated throughput performance of a physical host. This metric also helps us to understand other factors that may influence the aggregated performance.

c) *Network I/O per second* (*KB*/*s*). We measure the amount of network I/O traffic in KB per second, transferred to/from a web server during the corresponding workload.

4.2 Scheduling of cloud resources for CIA

Existing cloud resource scheduling strategies are mainly designed for Internet application, but CIA has a completely different set of application characteristics. First, the traditional CIA generally runs on a dedicated system or High Performance Computing (HPC) system, so its execution environment is dedicated or pre-configured with less dynamic resource scheduling. In the cloud environment of CIA, each layer has complex functions. For example, the IaaS layer is a virtual cluster with a group of specific constraints, the PaaS layer is a processing framework including computing scripts under its control, the SaaS layer is a particular application service, and also there are possibly combination of the above three layers, which is more complex. Therefore, the resource-scheduling problems CIA must face are more complex. Second, the goal of cloud resource-scheduling for CIA is to map the virtual machines in the cluster to the appropriate physical servers according to corresponding constraints, and enable multi-level synergistic scheduling between physical machines, multi-core processors and virtual machines. At the

Fig. 6 The comparison of IOPS for the random access pattern



same time, the cloud resource scheduling management needs to consider the lifecycle management of virtual machines, as well as resource reserves or queue sharing scheduling policies of virtual machines, which is more complex than traditional resource scheduling. Third, another goal of the cloud resource scheduling for CIA is to assign processing tasks to the appropriate computing nodes according to various characteristics of CIA's component, which also increases the difficulty of resource scheduling. Finally, CIA has an enormous number of components and a complex process flow, so the single reliability safeguard mechanism cannot adapt to the application characteristics of CIA. Therefore, it is necessary to establish a better reliability safeguard mechanism to ensure the high reliability of CIA.

To solve the resource schedule problems in cloud environments for CIA, we can gain inspiration from four aspects, the resource mapping strategy, the multi-layer scheduling model, the hybrid CPU/GPU scheduling algorithm, and the reliability guarantee mechanism. Figure 7 shows the entire technical framework.

1. *Multi-layer scheduling model for CIA*

The key to multi-layer scheduling for CIA is how to solve the coordination-scheduling problem between multi-layer resources. For example, the virtual cluster may require that execution rate of all virtual machines must be coordinated to be roughly the same. This scheduling demand not only includes the problem of virtual cluster scheduling, but also the problem of physical resources scheduling. The latter must be guaranteed by coordinated scheduling between virtual nodes and physical nodes. We suggest establishing the resource scheduling relationship model between virtual layers and physical layers, and design multi-layer scheduling protocols to ensure the harmonization of multi-layer scheduling.

2. *Resource mapping strategy for CIA*

The problem of resource mapping in cloud computing environments for CIA covers three aspects, the overall performance optimization, resource utilization maximization, and dynamic resource mapping to adapt the process complexity of CIA. In order to completely solve the resource-mapping problem and consider the three aspects above, we may adopt the following three steps. The first step is to analyze and refine the calculation and process features of CIA, and then evaluate the physical hardware computing resources. The second step is to establish the relationship model for different characteristics of computing tasks and computing environments. The third step is to create the resource mapping strategy for performance optimization of CIA. If cloud computing resources are abundant, in order to achieve the best overall performance of CIA, it is necessary to maximize the communication ability between virtual machines, especially to ensure low delay communication between virtual nodes. In a cloud environment,

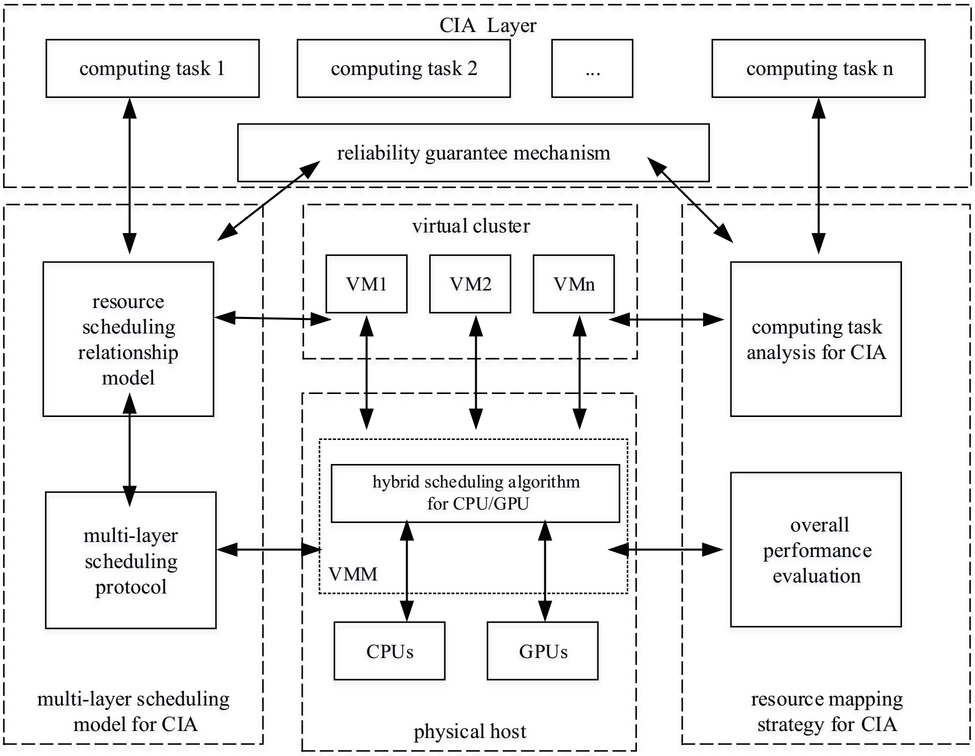

CIA Layer
computing task 1
computing task 2
...
computing task n
reliability guarantee mechanism
virtual cluster
VM1
VM2
VMn
resource scheduling relationship model
computing task analysis for CIA
hybrid scheduling algorithm for CPU/GPU
VMM
multi-layer scheduling protocol
overall performance evaluation
CPUs
GPUs
multi-layer scheduling model for CIA
physical host
resource mapping strategy for CIA

Fig. 7 Resources scheduling framework for CIA in cloud environment

there are many complex engineering tasks. If we only meet the best overall performance of each CIA one by one, many global resources of cloud computing will be wasted. The best way is to calculate the cost of each CIA, then adjust the overall performance of a single CIA through appropriate resource sharing and reuse, which may result in the overall maximum effectiveness of cloud computing resources.

3. *Hybrid scheduling algorithms for CPU/GPU*

In order to implement a fair and effective shared accelerator between multiple virtual machines, a feasible solution is to add a GPU in the VMM virtualization layer, which may provide the unified virtual accelerator interface for application layers. When application layers attempt to invoke the GPU interface, VMM will take over the control for the execution, which can enable the execution of multiple applications on the GPU in parallel. This method can make full use of the GPU parallel execution ability and implement hybrid scheduling that combines CPU and GPU.

We propose an improved virtualization framework for GPU/CPU, which is based on gVirtualS [42]. We introduce a new component called M-CUDA (Manager of Compute Unified Device Architecture) in VMM. This component can isolate and schedule resources of GPUs effectively for different virtual machines. The functions of M-CUDA include the following characteristics. a) *Dynamic scheduling*: when calculation task occupies the GPU resource for more than a certain idle time, M-CUDA will recycle the GPU resource. When the calculation task requests GPU resources again, M-CUDA allocates GPU resources for it. b) *Load balancing*: when local calculation pressure is too high, M-CUDA will adjust the computational load through dynamic scheduling to disperse the calculation load. c) *Failure recovery*: When the fault occurs, the calculation task is transferred to the new GPU resources available.

The computational efficiency of a GPU is proportional to the number of cores and its clock frequency. GPU global memory size is also an important factor for computational efficiency. If computation memory demand is greater than the GPU global memory size, the computation task is not completed at one time, which will introduce additional communication overhead. Therefore, the scheduling algorithm of GPU resources must consider three factors: the number of cores, the clock frequency and the size of global memory. To schedule GPU resources effectively, we set up a formula, as shown in the formula (1), to indicate the comprehensive evaluation of GPU resource loads.

$$L = \frac{\sum_{i=1}^{N}(S_i * C_i)}{a * G_c * \mathrm{F} + b * M} \tag{1}$$

In the formula, N indicates the number of all computation tasks running on the GPU, S_i indicates the scale of the *i*-th computation task, C_i indicates the computation complex of the *i*-th computation task, *Gc* indicates the number of cores of GPUs, F indicates the clock frequency of GPU, and M indicates the global memory of GPU. Finally, the symbol *a* indicates the impact factor of GPU computation ability and the symbol *b* indicates the impact factor of global memory. The value range of both *a* and *b* are between 0 and 1 and satisfies the condition $a + b = 1$. The values of *a* and *b* depend on the load property. If the load demands higher computation ability, then the value of *a* is bigger, and if the load demands higher global memory, then the value of *b* is bigger.

In order to verify the process of the value set of *a* and *b*, we assume a computation-intensive task, $C = A*B$, where *A*, *B* and *C* are $N*N$ matrices thus the elements of *C* are calculated as the following formula (2):

$$C_{ij} = \sum_{k=1}^{N} A_{ik} B_{kj} \tag{2}$$

We integrated the GPU scheduling mechanism on the Xen-based hypervisor and ran the above matrix computation task. Figure 8 shows the relationship of the running time of the task and the value of *a*.

We can see that when $a \approx 0.7$, the running time of the task is shortest. In order to compare the scheduling performance of local-GPU with that of M-CUDA, we ran the above matrix task on the same physical machine in both the single-task and multi-task modes. Figure 9 shows the comparison of the running time of a single-task mode for local-GPU and M-CUDA. We can see that the running time of local-GPU is less than that of M-CUDA. The main reason is that M-CUDA has internal communication and scheduling overhead. However, with the increase of the computation scale (the value of *N*), the running time of M-CUDA is almost the same as that of local-GPU. Figure 10 shows the comparison of finished tasks in multi-task mode for local-GPU and M-CUDA. We can see that the number of finished tasks is more than that of local-GPU under the premise that the number of submitted tasks is the same. In addition, through shared memory technology, the communication between M-CUDA and VMM also can be reduced, which may improve the performance of GPU/CPU greatly.

4. *Reliability guarantee mechanism for CIA*

CIA has a complex process flow, and the reliability of the whole process is necessary. An effective method

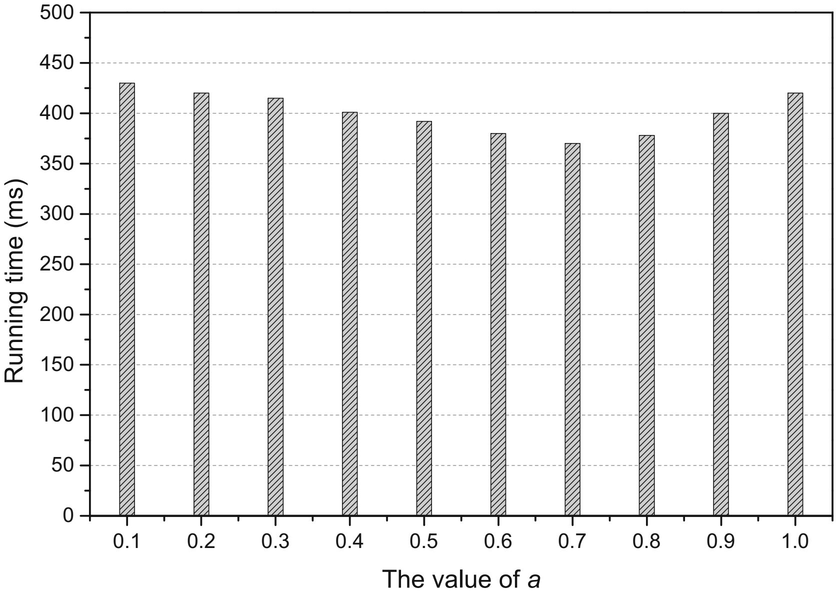


Fig. 8 Running time of tasks with different *a* and *b*

to guarantee CIA's reliability is a redundancy control service, also called synchronous hot standby. The various process stages of CIA involve various cloud services, such as IaaS, PaaS and SaaS. The reliability of PaaS and SaaS can be guaranteed using traditional parallel and distributed computing technologies, but the reliability of IaaS is not so simple, because it cannot be guaranteed by the reliability of each independent virtual machine. Redundancy control services create synchronous checkpoints of all virtual machines of CIA, used to find fault recovery points within the scope of the cluster. When a failure occurs, all of the virtual machines will be restored from the checkpoint, thus ensuring that the subsequent calculation is correct. The difficulty is how to determine the synchronization checkpoints of all virtual machines in a virtual cluster. A solution is to let virtual clusters enter a specific status, called "half synchronous consistent-coordination status", when it begins to synchronize checkpoints. The virtual machines that have entered this status cannot send communication data, but can receive communication data. Until all of the virtual machines have entered this status, they start to create consistent synchronous checkpoints. This solution can guarantee that the checkpoints of all virtual machines are consistent within the scope of the virtual cluster. The synchronization process may bring extra time overhead, but the influence on overall performance of CIA is very slight.

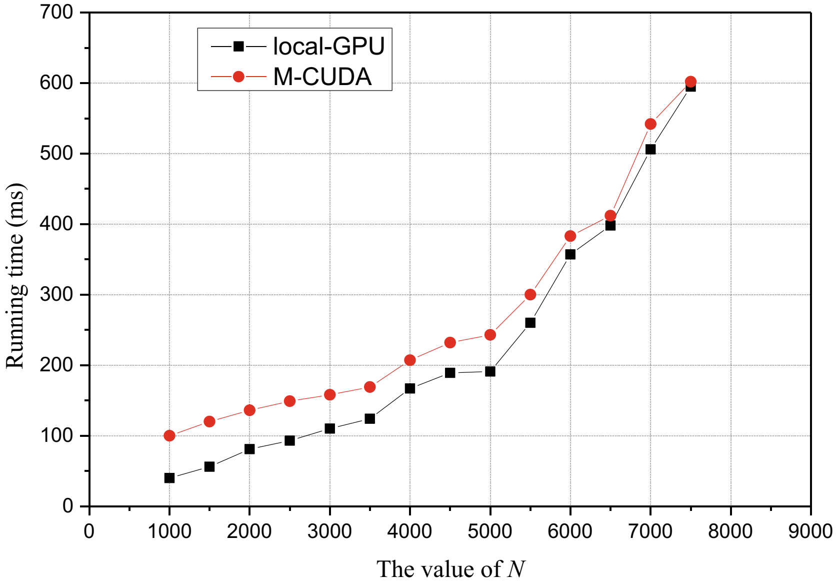
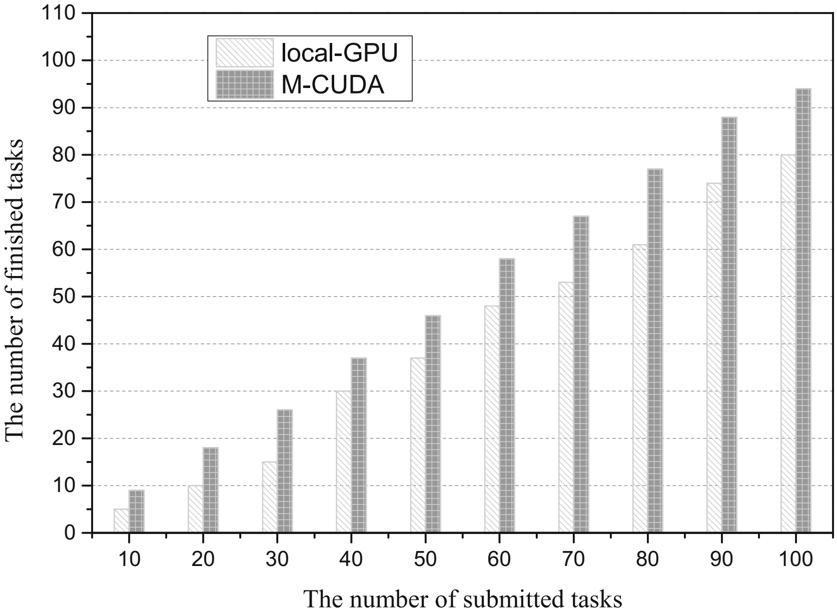


Fig. 9 Comparison of running time in single-task mode for local-GPU and M-CUDA

Fig. 10 Comparison of finished tasks in multi-task mode for local-GPU and M-CUDA



4.3 LCM

The CIA will give rise to the novel CPS platforms geared towards supporting collaborative industrial business processes and the associated business networks for all aspects of smart enterprises and smart product life cycles. With the support of CCPS, all the data during product life cycles can be automatically collected and forwarded to industrial cloud platforms using a wide range of communication technologies, which provides a huge possibility of realizing the LCM. Even so, LCM is still facing some issues and challenges, such as data mining targeting on CIA, response mechanism, and security and privacy.

1) *Data processing methods*

In a practical application, we are concerned that enterprises and individuals can rapidly extract the key information from massive industrial data to bring values or get better services. The current processing methods for massive industrial data include five features. a) *Bloom Filter*: A Bloom Filter consists of a series of Hash functions. b) *Hashing*: Hashing is a method that essentially transforms data into shorter fixed-length numerical values or index values. c) *Index*: Index is an effective method to reduce the expense of disk reading and writing, and improve insertion, deletion, modification, and query speeds. d) *Trial*: Trial is mainly applied to rapid retrieval and word frequency statistics. e) *Parallel Computing*: Parallel computing refers to simultaneously utilizing several computing resources to complete a computation task. The classic parallel computing models include MPI (Message Passing Interface), MapReduce, Dryad, etc. These methods have been proved effective in many applications. However, information of LCM for CIA is always partially known and partially unknown because of the complexity of CCPS and information confidentiality. We should introduce new approaches to analyze the massive industrial data. For example, we may consider using the grey systems-based techniques [43] for CIA to tackle uncertain information more effectively and efficiently.

2) *Response mechanisms*

The traditional response mechanisms between customers and service providers are realized by designing a reservation system, which has been widely used in healthcare and banking services. However, this approach has obvious deficiencies, such as lack of real-time interaction. The platform of CCPS can increase mutual understanding, but it results in new problems. On the one hand, customers could be reluctant to offer more information (e.g., device status) to service providers; they only hope for better services. On the other hand, service providers look forward to gathering all the data during product life cycles, so these are conflicting goals. If so, it is helpful for product-optimization design, and offers better services for customers. In order to achieve a good tradeoff, we must introduce the rational incentive mechanisms, such as a Mechanism of more Contributions and more Feedback Services (termed MCFS), to attract the prospective user's attention and participation.

5 Conclusions

With the advances in emerging technologies, it is feasible to seamlessly integrate CPS with cloud computing, which provides tremendous opportunities for CIA. In this article, we provided a brief review and outlook of the related background, and discussed the CCPS architecture and the enabling technologies for CIA. In particular, we dissected three key challenges (virtualized resource management techniques, scheduling of cloud resources for CIA, and LCM) and provided possible solutions to improve the performance and QoS of CCPS. The future work should include the following aspects: 1) establishment of a prototype for CCPS; 2) information exchange mechanism among the various devices; and 3) big data-based system optimization. We believe that CCPS for CIA will attract enormous attention and research effort in the near future.

Acknowledgements This work is partially supported by National Natural Science Foundations of China (Nos. 61572220, 61262013, 51575194, 61472283 and 61103185), the Fundamental Research Funds for the Central Universities (Nos. 2015ZZ079, 2013KJ034, 2100219043 and 1600219246), the Natural Science Foundation of Guangdong Province, China (No. 2015A030313746), the Fok Ying-Tong Education Foundation, China (Grant No. 142006), and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry.

References

1. Wang S, Wan J, Li D, Zhang C (2015) Implementing smart factory of industrie 4.0: an outlook. Int J Distrib Sensor Netw, 2015, Article ID 681806, 10 pages. DOI:10.1155/2015/681806.
2. Wan J, Zhang D, Zhao S, Yang L T, Lloret J (2014) Context-aware vehicular cyber-physical systems with cloud support: architecture, challenges and solutions. IEEE Commun Mag 52(8):106–113
3. Sridhar S, Hahn A, Govindarasu M (2012) Cyber-physical system security for the electric power grid. Proc IEEE 100(1):210–224
4. Banerjee A, Venkatasubramanian K, Mukherjee T, Gupta S (2012) Ensuring safety, security, and sustainability of mission-critical cyber-physical systems. Proc of the IEEE 100(1):283–299
5. Rajhans A, Bhave A, Ruchkin I, Krogh BH, Garlan D, Platzer A, Schmerl B (2014) Supporting Heterogeneity in Cyber-Physical Systems Architectures. IEEE Trans Autom Control 59(12):3178–3193
6. Derler P, Lee EA, Alberto SV (2012) Modeling cyber-physical systems. Proc IEEE 100(1):13–28
7. Chen F, Deng P, Wan J, Zhang D, Vasilakos A, Rong X (2015) Data Mining for the Internet of Things: Literature Review and Challenges. Int J Distrib Sensor Netw, 2015, Article ID 431047, 14 pages. DOI:10.1155/2015/431047.
8. Caliskan S, Rungger M, Majumdar R (2014) Towards robustness for cyber-physical systems. IEEE Trans Autom Control 59(12): 3151–3163
9. Wan J, Zhang D, Sun Y, Lin K, Zou C, Cai H (2014) VCMIA: a novel architecture for integrating vehicular cyber-physical systems and mobile cloud computing. Mobile Networks and Applications 19(2):153–160
10. Chen M, Zhang Y, Li Y, Mao S, Leung V (2015) EMC: emotion-aware mobile cloud computing in 5G. IEEE Netw 29(2):32–38
11. Rajkumar R (2012) A cyber–physical future. Proc IEEE 100(2): 1309–1312
12. Vamvoudakis KG, Hespanha JP, Sinopoli B, Mo Y (2014) Detection in adversarial environments. IEEE Trans Autom Control 59(12):3209–3223
13. Demirel B, Zou Z, Soldati P, Johansson M (2014) Modular design of jointly optimal controllers and forwarding policies for wireless control. IEEE Trans Autom Control 59(12):3252–3265
14. Trimpe S, D'Andrea R (2014) Event-based state estimation with variance-based triggering. IEEE Trans Autom Control 59(12): 3266–3281
15. Liu Q, Wan J, Zhou K (2014) Cloud manufacturing service system for industrial-cluster-oriented application. Journal Internet Technology 15(3):373–380
16. Zhang D, Wan J, Liu Q, Guan X, Liang X (2012) A taxonomy of agent technologies for ubiquitous computing environments. KSII Trans Internet Infor Syst 6(2):547–565
17. Wan J, Yan H, Li D, Zhou K, Zeng L (2013) Cyber-physical systems for optimal energy management scheme of autonomous electric vehicle. Comput J 56(8):947–956
18. Baliga J, Ayre RWA, Hinton K, Tucker RS (2011) Green cloud computing: balancing energy in processing, storage, and transport. Proc IEEE 99(1):149–167
19. Warneke D, Kao O (2011) Exploiting dynamic resource allocation for efficient parallel data processing in the Cloud. IEEE Trans Parallel Distrib Syst 22(6):985–997
20. Khazaei H, Mišić J, Mišić VB, Rashwand S (2013) Analysis of a pool management scheme for cloud computing centers. IEEE Trans Parallel Distrib Syst 24(5):849–861
21. Jain R, Paul S (2013) Network virtualization and software defined networking for cloud computing-a survey. IEEE Communications Managzine 2013:24–31
22. Xiao Z, Xiao Y (2013) Security and privacy in cloud computing. IEEE Commun Surv Tutorials 15(2):843–859
23. Drath R, Horch A (2014) Industrie 4. 0: Hit or Hype? IEEE Ind Electron Mag 8(2):56–58
24. Chen M, Wan J, González S, Liao X, Leung V (2014) A survey of recent developments in home M2M networks. IEEE Commun Surv Tutorials 16(1):98–114
25. Chen M, Zhang Y, Li Y, Hassan M, Alamri A (2015) AIWAC: affective interaction through wearable computing and cloud technology. IEEE Wirel Commun 22(1):20–27
26. Li D, Li F, Huang X (2010) A model based integration framework for computer numerical control system development. Robot Comput Integr Manuf 26(4):333–343
27. Z. Shu, D. Li, Y. Hu, F. Ye and J. Wan. From models to code: automatic development process for embedded control system. *Proc. of IEEE Int. Conf. on Network*, *Sensor and Control*, Shanya, China, pp. 660–665, 2008.
28. Wan J, Li D, Yan H, Zhang P (2010) Fuzzy feedback scheduling algorithm based on central processing unit utilization for a software-based computer numerical control system. Proc. Inst Mech Eng, Part B: J Eng Manufac, 224(7):1133–1143
29. J. Wan and D. Li. Fuzzy feedback scheduling algorithm based on output jitter in resource-constrained embedded systems. Proc. of Int. Conf. on Challenges in Environmental Science and Computer Engineering, Wuhan, China, pp. 457–460, Mar. 2010.
30. CIWA (Chinese Industrial Wireless Alliance). WIA-PA. [Online]. Available: http://www.industrialwireless.cn/.
31. J. Liu, Q. Wang, J. Wan and J. Xiong. Towards Real-time Indoor Localization in Wireless Sensor Networks. In *Proc. of the 12th IEEE Int. Conf. on Computer and Information Technology*, Chengdu, China, October, 2012, pp. 877-884.
32. Liu J, Wan J, Wang Q, Deng P, Zhou K, Qiao Y (2015) A survey on position-based routing for vehicular Ad hoc networks. Telecommun Syst. doi:10.1007/s11235-015-9979-7

33. Tian W, Zhao Y (Oct. 2014) Optimized cloud resource management and scheduling: theories and practices. Elsevier/Morgan Kaufmann, Publisher
34. Chen M, Hao Y, Li Y, Lai C, Wu D (2015) On the computation offloading at Ad Hoc cloudlet: architecture and service models. IEEE Commun Mag 53(6):18–24
35. H. Suo, Z. Liu, J. Wan and K. Zhou. Security and Privacy in Mobile Cloud Computing. *Proc. of the 9th IEEE Int. Wireless Communications and Mobile Computing Conf.*, Cagliari, Italy, Jul. 2013.
36. Chen M, Mao S, Liu Y (2014) Big data: a survey. Mobile Networks and Applications 19(2):171–209
37. Yuan W, Deng P, Taleb T, Wan J, Bi C (2015) An Unlicensed Taxi Identification Model based on Big Data Analysis. IEEE Trans Intell Transp Syst. doi:10.1109/TITS.2015.2498180
38. I. King, J. Li and K. T. Chan. A brief survey of computational approaches in social computing. Neural networks, *IEEE-INNS-ENNS International Joint Conference on*, pp. 1625–1632, Jun. 2009.
39. FaceBook, [Online]. Available: http://www.facebook.com.
40. Shen W, Hao Q, Mak H, Neelamkavil J, Xie H, Dickinson J, Thomas R, Pardasani A, Xue H (2010) Systems integration and collaboration in architecture, engineering, construction, and facilities management: A review. Adv Eng Inform 24(2):196–207
41. H. Yan, J. Wan, D. Li, Y. Tu and P. Zhang. Code sign of Networked Control Systems: A Review from Different Perspectives. *Proc. of IEEE Intl Conf. on Cyber Technology in Automation*, *Control*, *and Intelligent Systems*, Kunming, China, pp. 84–90, Mar. 2011.
42. Giunta G, Montella R, Agrillo G, Coviello GA (2010) GPGPU transparent virtualization component for high performance computing clouds. In: In Euro-Par 2010-Parallel Processing. Springer Berlin Heidelberg, pp. 379–391
43. Liu S (2006) and Y. Li. Grey Information, Theory and Practical Applications. Springer Science & Business Media

STATEMENT

This work is based in part on our previous paper titled “An Insight into Cloud-enabled Complex Industrial Applications,” Presented at the fourth IEEE/CIC International Conference on Communications in China (ICCC 2015), Shenzhen, China, November, 2015.