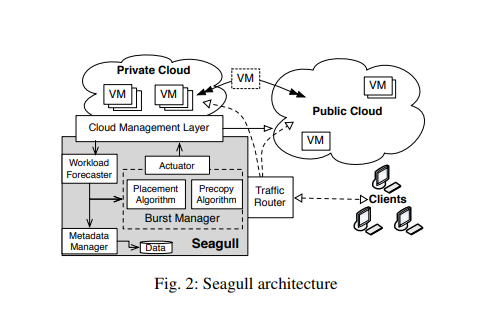
Abstract:

The cloud based delivery model for IT resources is revolutionizing the IT industry.[1] The practice of computing across two or more data centers separated by the Internet is growing in popularity.[8] Cloud computing is changing the infrastructure upon which scientific computing depends from supercomputers and distributed computing clusters to a more elastic cloud-based structure.[6]The high cost of provisioning resources to meet peak application demands has led to the widespread adoption of pay-as-you-go cloud computing services to handle workload fluctuations.[2] For many organizations, one attractive use of cloud resources can be through what is referred to as cloud bursting or the hybrid cloud.[3] While cloud-bursting is addressing this process of scaling up and down across data centers (i.e. between private and public clouds), offering service level guarantees, is a challenge for inter-cloud computation, particularly for best-effort traffic and large files.[8] Enterprises with existing IT infrastructure are beginning to employ a hybrid cloud model where the enterprise uses its own private resources for the majority of its computing, but then “bursts” into the cloud when local resources are insufficient. However, current approaches to cloud bursting cannot be effectively automated because they heavily rely on system administrator knowledge to make decisions.[4] However from the SMEs perspective, there is still a barrier to Cloud adoption, being the need of integrating current internal infrastructure with Clouds. They need strategies for growing IT infrastructure from inside and selectively migrate IT services to external Clouds in a way that enterprises benefit from both Cloud infrastructure’s flexibility and agility as well as lower costs.[5] The transformation of scientific computing infrastructure from traditional supercomputers and clusters to more flexible cloud-based systems. It highlights Fermi National Accelerator Laboratory's development of its private cloud, FermiCloud, and the potential for cloud bursting to accommodate increased computation demands.[6] The increasing popularity of computing across multiple data centers over the Internet to meet growing computational demands. It addresses challenges in achieving service level guarantees for inter-cloud computation, particularly for real-time, data-intensive tasks involving image and document processing. The proposed solution involves autonomic cloud-bursting schedulers that adapt to workload variations, bandwidth constraints, and available resources to provide probabilistic service level guarantees, focusing on speeding up processing and preserving job order in a document processing workload.[8] The trend of companies, including SMEs, embracing private cloud infrastructure to facilitate IT service outsourcing. However, it highlights the challenge of seamless extensibility via cloud bursting in existing software solutions, which requires users to prepare images for each infrastructure. The paper introduces a method using nested virtualization to address these issues, reducing complexity in the cloud bursting process. Performance tests show that this approach incurs a 5-10% overhead on deployment time but enhances seamless extensibility.[9]

Introduction :

Cloud computing as a highly scalable and flexible technology utilizing utility computing and service-oriented architecture. It discusses different deployment models (public, private, hybrid, internal, and external clouds) and architectural service layers (SaaS, PaaS, and IaaS). Privacy and security concerns, as well as a lack of enterprise-level adoption, are highlighted as challenges. The paper then outlines the security issues related to cloud services and the need for more mature delivery and operational models. The passage also mentions the OPTIMIS project, which addresses these issues, and provides an overview of the paper's structure, which includes discussions on cloud bursting, brokerage service models, and specific use cases for storage and compute services. the critical business drivers for cloud computing adoption, including cost reduction, increased agility, faster ROI, and enhanced business continuity. It acknowledges the significance of addressing privacy and security concerns to encourage broader enterprise adoption while noting the complexity of handling data privacy in multi-tenant cloud environments.[1] A system called Seagull that addresses the challenges of managing dynamic workloads in enterprise applications through a hybrid "cloud bursting" approach. Seagull automates the process of detecting resource overloads in a private cloud, decides which applications can be efficiently moved to a public cloud to minimize costs, and performs migrations to expand capacity. It supports both horizontally and vertically scalable applications and allows for the proactive replication of overload-prone applications to reduce migration time. The system is evaluated using a Xen-based local data center and Amazon EC2 cloud platform, demonstrating significant cost savings and faster application migrations. This work makes contributions in terms of efficient application placement algorithms, precopying strategies, and practical implementation for optimizing cloud bursting in large-scale data centers.[2]



cloud computing in scientific and data-intensive computing, particularly within the realm of high-performance computing (HPC). Cloud providers like Amazon have recognized the demand for HPC-specific services such as Cluster Compute Instances. The concept of "cloud bursting" is introduced, allowing organizations to seamlessly augment local resources with cloud capabilities to address spikes in computing or data needs. a middleware designed for hybrid data analysis, where data is distributed across local clusters and the cloud. This middleware supports Map-Reduce-style APIs, optimizing resource utilization and data distribution.[3] the use of cloud computing for handling dynamic workloads in enterprise applications. Many organizations have invested in local IT infrastructure but face occasional resource shortages during workload spikes. To address this, a hybrid model called "cloud bursting" is introduced, allowing the organization to utilize cloud resources during peaks. A cost-saving scenario is illustrated, showing that a hybrid approach can reduce expenses by up to 29%. However, existing virtualization tools lack automation and decision-making capabilities for efficient cloud bursting. The Seagull system is introduced to address these challenges. It automates the decision of which applications to move to the cloud at the lowest cost and performs migrations to expand capacity in response to workload spikes. [4] the rise of cloud computing as a flexible and cost-effective solution for IT resource provisioning, offering scalability and on-demand capacity. Major IT players such as Google, Amazon, Microsoft, and Oracle have embraced cloud computing, leading to the concept of the "Public Cloud" where applications and services are accessible on a pay-per-use basis. While the benefits of cloud computing are clear, there are concerns regarding performance, security, and legal issues for certain enterprise applications. The text suggests that a hybrid approach, combining in-house infrastructure with public cloud resources, is a practical solution. This combination, known as a hybrid cloud, allows enterprises to maintain critical operations in-house while leveraging the scalability and flexibility of public clouds. To realize this vision, the text emphasizes the importance of software middleware and scheduling policies for seamlessly provisioning resources from both local and public cloud infrastructures. The Aneka middleware platform is introduced as a solution that enables the dynamic expansion of compute-intensive applications in hybrid clouds, ensuring a balance between privacy, performance, and cost savings. The chapter also presents and evaluates scheduling approaches to optimize resource management for enterprises.[5] The challenges in big data analytics within private clouds due to exploding data sizes and increasing complexity. Private clouds often struggle to meet the scalability and deadline requirements. Cloud bursting, a form of hybrid cloud computing, has gained popularity to overcome these limitations by temporarily enhancing on-premise resources with off-premise resources from public cloud providers during peak utilization. The main challenge in enabling cloud bursting for large-scale big data analytics is the bottleneck caused by slow network links between on-premise and off-premise resources. This bottleneck necessitates new "hybrid cloud big data analytics" approaches. Iterative applications are well-suited for hybrid clouds, given their reuse of invariant input data. The text describes two techniques to accelerate such applications, including extended off-premise HDFS storage and locality-enforced scheduling, showing significant speed-up compared to default Hadoop implementations.[7] Cloud computing is a dominant IT paradigm, offering flexible pay-per-use services based on virtualization technology. It encompasses various service and deployment models, with a focus on Infrastructure-as-a-Service (IaaS) clouds providing fundamental IT resources like networking and computation. Amazon's introduction of Elastic Cloud Computing (EC2) in 2006 marked the inception of public IaaS providers, followed by many others and the emergence of private cloud solutions. Private IaaS clouds enable the integration of private and public clouds, often via cloud bursting, allowing companies to extend their infrastructure dynamically with third-party resources. [9] PaaS (Platform as a Service) for cloud applications and highlights the need for Service Level Agreements (SLAs) that go beyond resource availability, addressing meaningful QoS properties such as response time and throughput. The paper introduces Meryn, a cloud bursting PaaS system architecture, designed to optimize provider profit while remaining extensible for various programming frameworks. Meryn uses a decentralized approach for resource distribution and incorporates a profit optimization policy. The prototype of this system demonstrates its effectiveness in maximizing provider profit and minimizing the use of public cloud resources, outperforming basic approaches. The paper is organized to discuss related work, present the architecture, explain the profit optimization policy, and provide experimental results.[13]

Literature of exist Research :

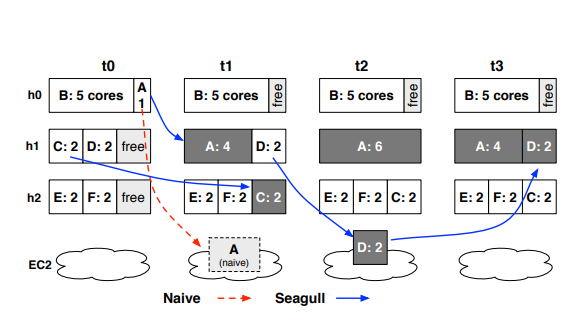
The adoption of cloud bursting can lead to substantial cost savings for enterprises. Instead of provisioning all resources locally, a hybrid approach of owning some resources locally and renting additional servers from the cloud during peak periods can yield significant savings. For example, a business requiring ten servers two days a week can save 27% of its costs through a hybrid approach. This concept has led to the development of hybrid cloud management tools by software vendors and providers like Amazon, VMware, and Rackspace. While existing solutions focus on resource migration at coarse time scales, the goal of this work is to enable agile cloud bursting that responds to workload spikes within hours or minutes. However, challenges remain in automating high-level policy decisions for migration in private cloud data centers with numerous applications.[2] scenarios where data processing in a hybrid cloud is desirable and the need for a supporting framework. The proposed framework is akin to Map-Reduce, enabling transparent remote data analysis, even when data and processing resources are geographically distributed, including cloud components. Unlike traditional solutions, it allows for the separation of data and processing, relieving developers from managing data movements. The paper focuses on scenarios where a portion of data is stored locally and the rest in the cloud, with processing resources available at both ends, accommodating variations in data and processing availability. The implementation and evaluation are specific to Amazon Web Services (AWS) but have broader applicability to pay-as-you-go cloud providers.[3] The current limitations in cloud technologies and introduces the EU FP7OPTIMIS project, which aims to optimize the entire cloud service life cycle. The project seeks to create an open, scalable, and dependable cloud service ecosystem that improves the delivery of IT services, with a focus on automation, security, and sustainability. It goes beyond existing solutions that primarily address infrastructure-related issues and emphasizes higher-level concerns, such as cloud bursting and brokerage. The paper particularly focuses on the security aspects of cloud bursting and brokerage, while acknowledging that issues like abstracting heterogeneity and cloud-based application development are beyond its scope.[1] Cloud services are experiencing rapid growth in enterprise adoption, with a global market predicted to reach over $140 billion in 2014. Researchers from various fields are migrating their applications to computer clouds due to the "pay-as-you-go" model and service-oriented approach. Private clouds, however, may have resource limitations for large scientific applications, leading to the integration of commercial clouds into a hybrid cloud setup. One successful example is the STAR project, which collaborated with the Nimbus team to migrate experiments to a computer cloud. Many researchers have contributed to building cloud federations, such as the Cross-Cloud Federation Manager (CCFM) and Cloud Scheduler, to facilitate cloud bursting and dynamic scientific applications, making cloud locations transparent to users and improving performance while reducing operational costs. the performance modeling of MapReduce applications, particularly focusing on iterative ones in the context of hybrid cloud bursting. While previous studies have explored MapReduce on single cloud platforms, issues like storage elasticity and data shuffling are critical for iterative applications. Specific optimization techniques for iterative MapReduce have been proposed in the past, but they can complement the techniques discussed in the paper. Performance and cost prediction within a single data center have also been studied, involving aspects such as storage performance, scheduling, resource provisioning, and job estimation. Existing models and frameworks are suited to single-site deployments and don't fully address the challenges associated with hybrid cloud bursting and iterative MapReduce applications. The paper introduces a novel approach that considers the unique challenges of performance modeling when dealing with both iterative MapReduce applications and hybrid cloud bursting, making it a pioneering effort in this context.[6] Evidence for the feasibility of real-time cloud bursting, particularly with the transfer of extremely large workloads over the regular Internet. It focuses on delivering service level guarantees for computational workloads involving documents and images. While some research has explored efficient brokering strategies and scheduling techniques, more investigation is needed, especially in the context of inter-cloud processing. The proposed slackness constraints offer a novel way to balance chronological priority with traditional service level agreements, making cloud bursting effective for small and medium organizations. The techniques and findings presented are applicable to various domains, including academic computing environments. The paper addresses key questions about determining when, where, and how much to burst workloads for optimizing downstream service level attributes, highlighting the opportunistic and autonomic nature of their schedulers.[8] Recent efforts in cloud bursting techniques and procedures have addressed data-intensive computations, batch job scheduling, and concepts such as cloud aggregation and brokerage. While these endeavors have shown promise, they often overlook challenges such as adapting applications for different IaaS clouds, administrative domain disparities in public clouds, and the need for live VM instance migration. Potential solutions like HVX for easy multi-VM application migration and Xen-blanket for similar functionality, though limited to the Xen hypervisor, aim to overcome some of these issues, ensuring flexibility and reliability in cloud infrastructures.[9] Efforts to enhance MapReduce frameworks in hybrid environments have seen significant research. HybridMR introduces a hybrid grid with voluntary computing nodes, while HadoopDB combines parallel databases with Hadoop for improved performance. Other studies employ GPUs for MapReduce acceleration in clusters. Bicer addresses data analysis in hybrid clouds where data is pre-partitioned, while elasticity techniques optimize cost-performance trade-offs in cloud environments. Weak links between on-premise and off-premise resources have been recognized as an issue impacting performance, particularly in Hadoop applications. Numerous studies explore scheduling strategies for hybrid cloud-based MapReduce applications, focusing on factors such as data migration and scheduling strategies to mitigate the impact of weak links on performance.[10] The proliferation of both commercial and open-source PaaS (Platform as a Service) systems, along with research PaaS platforms, has expanded the landscape of cloud application hosting. However, only a handful of these platforms support hosting applications on multiple IaaS (Infrastructure as a Service) clouds, and even fewer provide automated resource selection policies or profit optimization for service providers. In addressing optimization challenges, prior research has explored different facets. One study focuses on time and cost-sensitive execution of data-intensive applications in hybrid clouds, considering cost and time constraints. Another approach suggests a decentralized economic strategy to dynamically adjust cloud resources for web service applications to meet performance and availability SLAs while minimizing operational costs. An SLA-based resource allocation problem for multi-tier applications in cloud computing is tackled to optimize total profit, factoring in operational costs. Yet another research proposes a resource allocation algorithm for SaaS providers to minimize infrastructure cost and SLA violations.[12] Numerous PaaS platforms, including both commercial and open-source options, have emerged, but few of them support hosting applications across multiple IaaS clouds. Moreover, even those that do seldom offer automated resource selection policies or profit optimization for service providers. Prior research in this domain has delved into optimization challenges, such as balancing cost and time constraints for data-intensive applications in hybrid clouds and adopting decentralized economic strategies for dynamic resource adjustment in web service applications to meet SLAs while reducing operational costs. Additionally, studies have tackled SLA-based resource allocation problems, aiming to optimize profits for multi-tier applications while factoring in operational expenses. Another line of research introduces resource allocation algorithms for SaaS providers to minimize infrastructure costs and mitigate SLA violations.[13] The passage discusses pilot-based Workload Management Systems (WMS) used in grid and cloud computing environments, with a focus on HTCondor and GlideinWMS. HTCondor, originally designed for harnessing idle workstation cycles, has evolved into a significant resource management system. It comprises several components, including a central manager, startd for managing compute resources, schedd for maintaining job queues, and a negotiator for matching resources to user jobs. HTCondor's asynchronous communication allows for efficient resource allocation based on user priorities. GlideinWMS, built on top of HTCondor, is a pilot-based WMS. It consists of the GlideinWMS Provisioning Service, comprising the Glidein factory and VO Frontend services, and the Glidein Service that manages the environment on worker nodes for executing user jobs. This architecture enables better scalability and operational division between grid debugging and policy management. In essence, these systems provide efficient resource management and allocation for applications in grid and cloud computing environments.

Methods :

We have implemented Seagull’s placement and precopying algorithms as modules that extend the OpenNebula cloud management software. We have created a private cloud environment on a lab cluster using OpenNebula over the Xen-hypervisor and used Amazon EC2 as our public cloud. We use three applications, TPC-W, Wikibooks and CloudStone for our evaluation. We have created private-cloud as well as public-cloud appliances for each of these three applications and their respective client applications. An appliance instance will create the virtual machine(s) which house the complete application. We warm up each application, using its clients, for two minutes before collecting data.

**Cloud Bursting Time** The total time to perform a cloud burst can be decomposed into three major parts: copying data to the cloud, preparing an application image, and booting up the virtual machine. To measure each of these components, we migrate a virtual machine running the CloudStone application with a disk-state size of 5GB.

**Placement Algorithm** In this experiment, we analyze the placement efficiency of Seagull compared to a naive algorithm (which always cloudbursts the overloaded application), in a small scenario that demonstrates the intuition behind Seagull’s decision making.

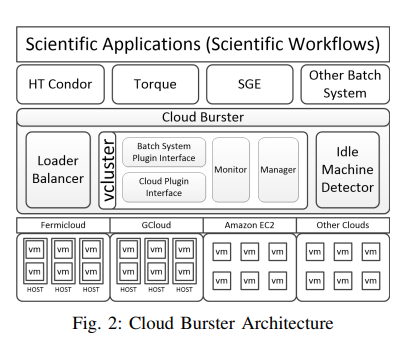


Seagull optimizes workload spikes by reshuffling local resources and efficiently moving applications between the data center and the cloud, saving costs. In contrast, the naive algorithm lacks this optimization, leading to higher cloud expenses and inefficient resource usage.

**An Efficient Greedy Bursting Algorithm**: Seagull's greedy algorithm prioritizes local resource utilization and cost-effective migration strategies during workload spikes. It first attempts to address overloads locally, using a sorted list of servers to find free capacity or perform VM repacking. If that's insufficient, it moves applications to the cloud, favoring those that free up the most local resources relative to their cloud cost, while considering resource requirements and cost of all affected applications on a host. This approach aims to minimize cloud usage and reduce overall expenses.

We define the cost of bursting an application, say A, that is composed of n VMs using:

CostA= Sum(from j=1 to n) of CostAjkl , Costijkl = Tijkl + (Rijkl ∗ τ ) + (Sijkl ∗ months(τ ))

**Cloud Bursting Architecture:** The cloud burster middleware facilitates automatic cloud bursting and aims to offer users an elastic virtual cluster for high-performance execution of scientific applications. It achieves this by aggregating resources from multiple IaaS providers into a unified resource pool, consisting of three key components: (i)-Vcluster, (ii)-Idle machine detector, (iii)-Load balancer.

**The novel cloud bursting technique:**

Principles: The authors emphasize the need for seamless cloud bursting, where (i) applications function the same on public clouds as on private ones without VM image modifications, (ii) users perceive no difference between the two environments, and (iii) live migration supports continuous service operation.

Design: They achieve this by using virtual private networking (VPN) and nested virtualization technology, creating a common networking environment for VM instances in both private and public clouds. Nested virtualization bridges differences between hypervisors and image formats among various IaaS solutions.

Cloud Bursting Process: Traditional cloud bursting involves users adapting their applications and images for both private and public clouds. In contrast, the novel technique launches virtual compute nodes on target clouds via VPN connections, making bursted VM instances nested on the target clouds. This approach doesn't require application adaptation and maintains the administrative domain. The primary focus is on extending computing capacity, but the same principles can be applied to networking and storage capacity expansion.

Research Gap :

Implementing steps to ensure SLA and security requirements match user needs and exploring cloud bursting and cloud aggregation models.[1] automating the cloud bursting process, addressing challenges like initiating cloud bursts without accurate workload predictions, standardizing cloud interfaces for live WAN migration, and integrating Seagull's placement algorithm with business policies.[4] , current approaches to cloud bursting cannot be effectively automated because they heavily rely on system administrator knowledge to make decisions [4]include extending these scheduler techniques to various job classes and exploring advanced models to handle Internet transience. Additionally, modulating job chunking based on queue position and conducting experiments with more resources are planned.[8] will involve extending the c.b. technique and concentrating on networking and storage capacity bursting.[8]

Conclusion :

The cloud delivery model is seen as a solution for complex IT systems, but the lack of mature operational models like cloud bursting and cloud brokerage hinders its adoption. This paper introduces concepts like cloud bursting, cloud brokerage, and cloud aggregation, focusing on the capability requirements.[1] "Seagull," a cloud bursting system that automates the efficient migration of applications from a private data center to the public cloud. It uses selective precopying to reduce migration time, enabling agile resource provisioning and cost minimization. Seagull's placement algorithm considers local re-consolidation opportunities and application costs, resulting in a 45% reduction in cloud bursting costs during data center overload, with minimal performance overhead.[2] a framework for data-intensive computing in hybrid clouds, combining local and cloud resources efficiently. Evaluation reveals that cloud bursting offers flexibility without significant performance degradation. Inter-cluster communication overhead is low, the middleware balances computation effectively, and remote data retrieval overheads have a modest impact on most applications, supporting the effectiveness of this approach.[3] "Seagull," a cloud bursting system that efficiently moves applications from a private data center to the public cloud when local resources are overloaded. This enables agile resource provisioning, optimizing local resource utilization while keeping cloud expenses minimal.[4] cloud bursting has overlooked workloads with transit times to external clouds similar to processing times, particularly due to large file sizes and bandwidth limitations. The study focuses on enabling small organizations to use cloud bursting effectively, emphasizing the need for sophisticated schedulers to adapt to workload dynamics. Three scheduler types are proposed, addressing large job cloud bursting and respecting queue constraints and service level agreements.[8] introduces cloud bursting techniques and related works, presenting a novel method using nested virtualization and advanced networking. Evaluation indicates that seamless cloud bursting results in a 5-10% increase in deployment time.[9]

References :