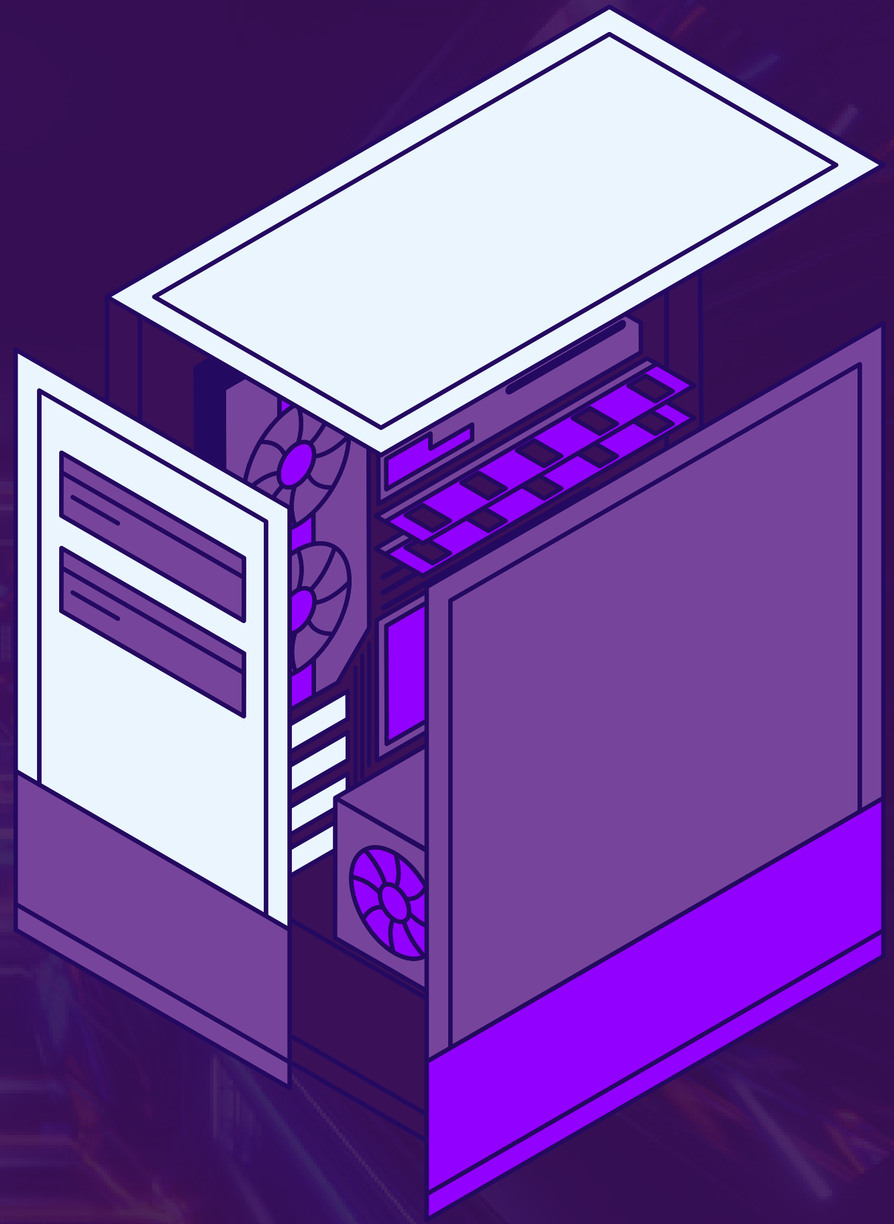
AI-DRIVEN RESOURCE MANAGEMENT IN CLOUD COMPUTING

Mete Harun Akcay, Masa Cirkovic, Mahira Joytu, Ahmad Alkhaldi, Alexis Gbeckor-Kove

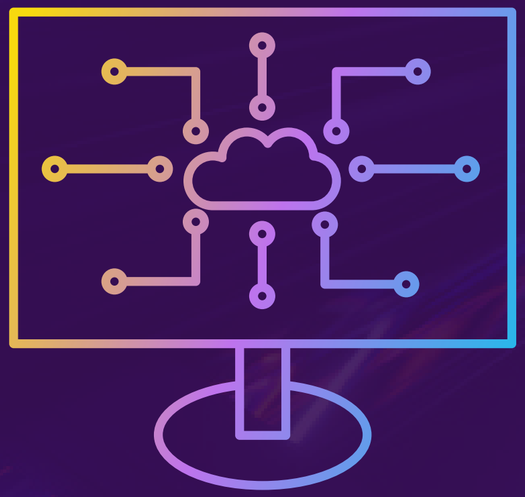
AGENDA

- INTRODUCTION
- TRADITIONAL RESOURCE MANAGEMENT
- CHALLENGES
- AI FOR RESOURCE MANAGEMENT
- CONCLUSION

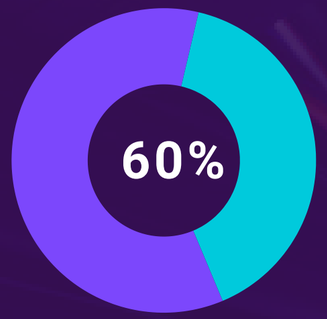


INTRODUCTION

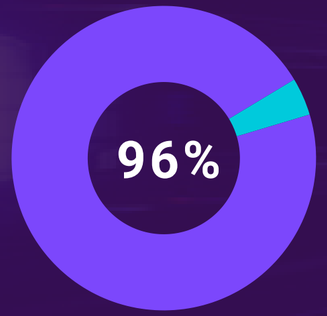
Cloud Computing: “Delivery of computing services - including servers, storage, databases, networking, software, analytics, and intelligence - over the internet to offer faster innovation, flexible resources, and economies of scale.”



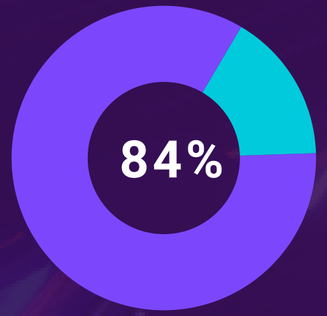
Cloud computing service models are based on the concept of sharing on-demand computing resources over the internet. Companies pay to access a virtual pool of shared resources which are located on remote servers that are owned and managed by service providers.



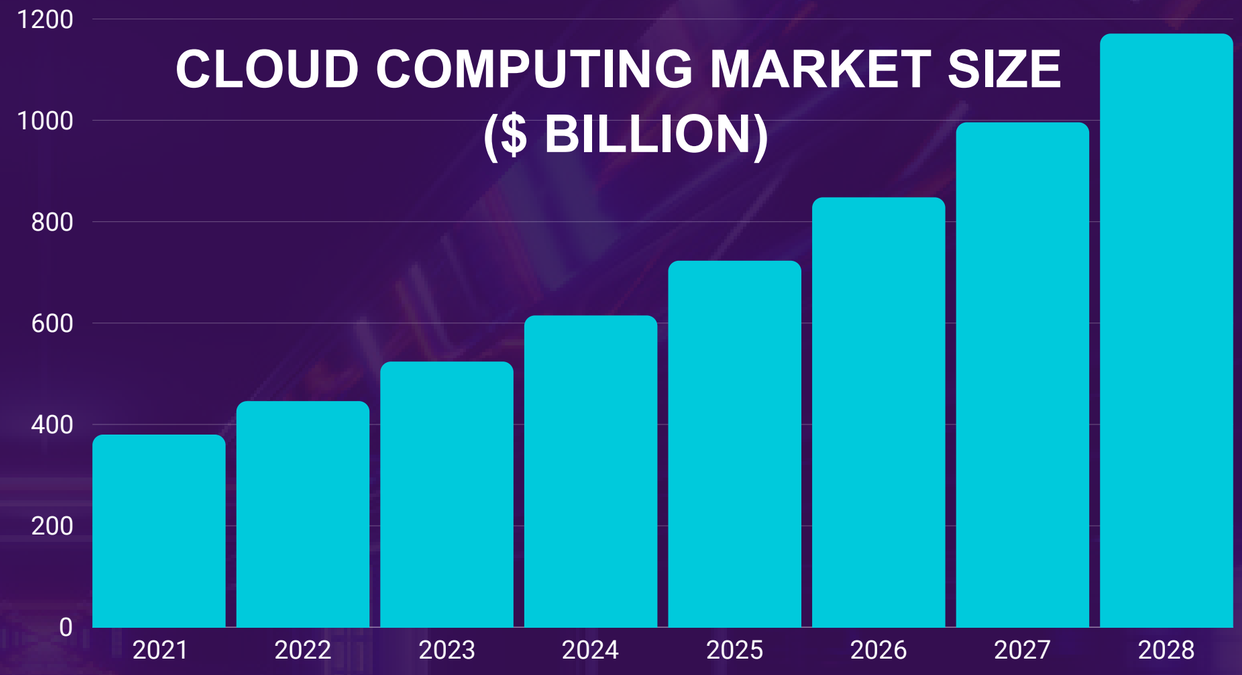
60% of the world’s corporate data is stored in the cloud.



96% of the companies use the public cloud.

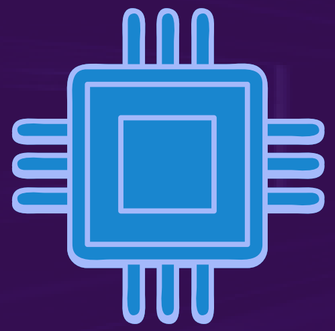


84% of the companies use the private cloud.



TRADITIONAL RESOURCE MANAGEMENT

Methods and strategies employed to allocate, optimize, and manage hardware and software resources. Goal is to ensure **efficient** and **balanced** use of resources, while maintaining **service quality**, minimizing **operational costs**, and avoiding **over-provisioning** or **under-provisioning**.



Computing resources

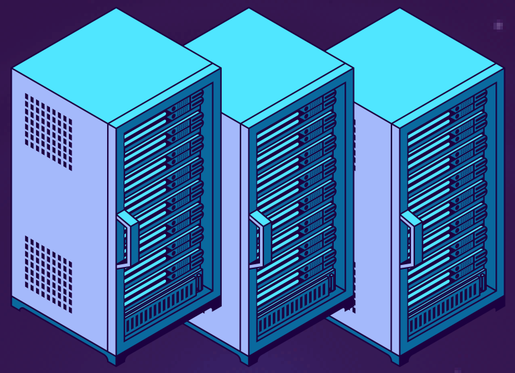
Storage resources

Network resources

KEY ASPECTS

Manual Resource Provisioning

- Static Allocation - allocated to applications based on peak load estimates. Manually assigned resources
- Predefined Thresholds - thresholds for resource usage. Administrators must scale up or down manually when thresholds are breached
- Suboptimal Utilization - no real-time optimisation, over-provisioning, under-utilization, increased costs



Capacity Planning

- Over-Provisioning - resource provisioning involves over-provisioning to ensure that peak demands are met, resulting in idle resources, inefficient utilization
- Under-Provisioning - insufficient resources for demands, performance degradation

KEY ASPECTS

Scheduling

- Job Scheduling - jobs scheduled based on static policies like FCFS, priority, round robin, static
- Fixed Resource Pools - resources devided into pools, distributed according to job req. Can result in time out
- Priority-Based Allocation - higher-priority jobs receive more resources, may lead to delays for less urgent tasks

Monitoring and Reporting

- Manual Monitoring - tracking resource consumption. Alerts are trigger based on static thresholds
- Limited Automation - monitoring and scaling requrire human intervention. Issues detected after impacting performance

KEY ASPECTS

Energy Efficiency

- High Power Consumption - lacks mechanisms to optimize power usage. Servers may remain powered on even when underutilized or idle, resulting in unnecessary energy consumption. Every under-utilization of active resources leads to energy waste.

Cost Management

- Capital Expenditure (CapEx) - involves significant upfront investments in hardware and infrastructure. This leads to high CapEx, and organizations may over-invest in resources that are rarely fully utilized.
- Operational Expenditure (OpEx) - ongoing operational costs associated with maintaining and managing the infrastructure, including staffing, power, cooling, and physical maintenance.

KEY ASPECTS

Load Balancing

- Simple Distribution Methods - predefined algorithms to distribute traffic across resources. Methods include round-robin, least connections, weighted round-robin
- Lack of Real-Time Adjustment - minimal real-time feedback to rebalance workload
- Manual Configuration - setting up algorithms, determining which servers are included

Manual Scalability

- Vertical Scaling (Scaling Up) - adding more resources to an existing server or VM to handle increased demand. Can can be expensive since over-provisioning may lead to resource waste.
- Horizontal Scaling (Scaling Out) - adding more servers or instances to the system to distribute the load. Requires manual intervention to add new servers to the resource pool. Configuring load balancers requires manual adjustments.

CHALLENGES IN RESOURCE MANAGEMENT

Scalability

- **Complexity of Scaling Resources -** Traditional resource management systems lack the dynamic and real-time adjustment capabilities required for unpredictable workloads
- **Integration of IoT and Edge Computing -** presents significant challenges due to the diverse and distributed nature of IoT devices

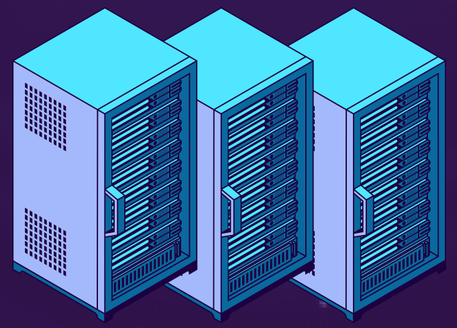
Cost Efficiency

- **Over- and Under-Provisioning -** Difficulty in accurately forecasting demand, leading to either over-provisioning (wasting resources) or under-provisioning (causing performance degradation)
- **Operational Overhead and Human Intervention -** Manual resource management methods increase operational overhead

CHALLENGES IN RESOURCE MANAGEMENT

Resource Allocation

- **Dynamic Workloads -** Difficulty in managing highly dynamic workloads that fluctuate unpredictably
- **Heterogeneity of Cloud Resources -** Managing heterogeneous resources within cloud environments, which include different types of computing power, storage, and bandwidth
- **Latency and Network Bottlenecks-** Poses significant challenge when integrating IoT and edge computing with cloud systems



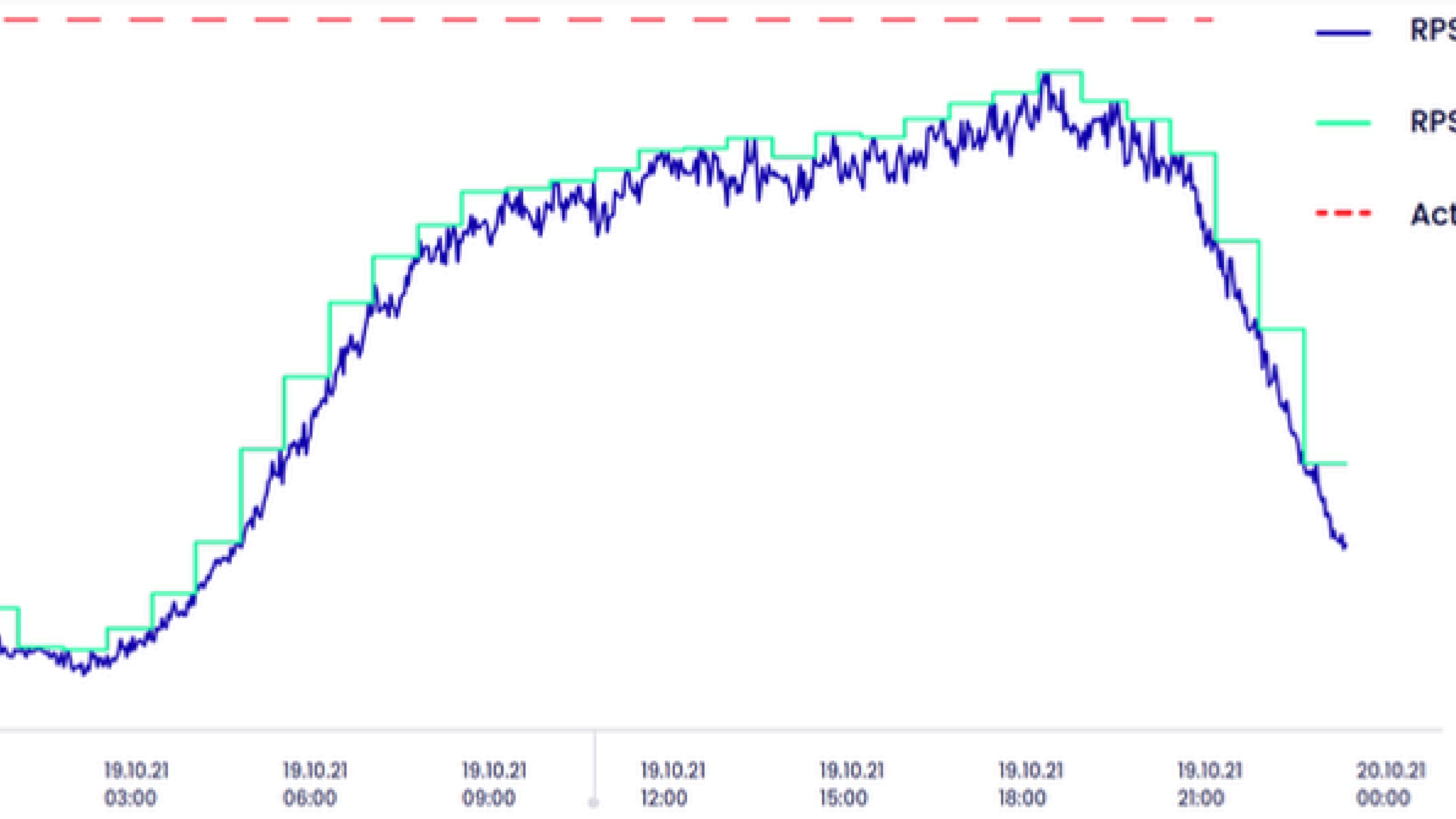
System Performance

- **Performance Degradation Due to Resource Mismanagement -** Traditional resource management approaches can lead to performance degradation due to the static allocation of resources
- **Security Concerns -** Performance can be affected by security challenges
- **Trade-off Between Performance and Cost -** There is often a trade-off between achieving high system performance and maintaining cost efficiency

AI FOR RESOURCE MANAGEMENT

AI-Driven Computing Resource Management

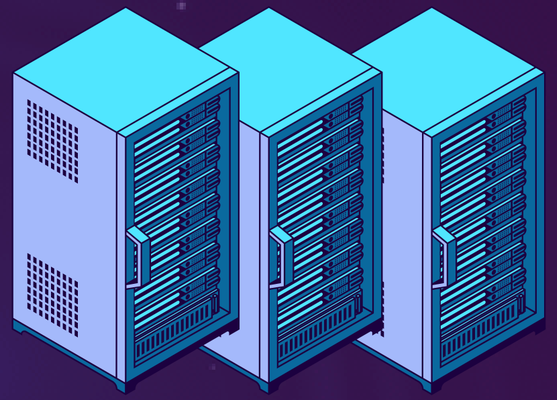
- Predictive Autoscaling - forecasts demands based on time-series models like LSTM and ARIMA to scale resources proactively.
- Traffic Pattern Analysis - Predicts future network usage based on historical data
- Dynamic Workload Scheduling- •Schedules tasks based on real-time demand and costs.

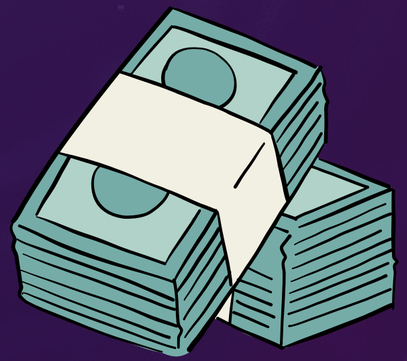

RPS
RPS
Act
19.10.21
03:00
19.10.21
06:00
19.10.21
09:00
19.10.21
12:00
19.10.21
15:00
19.10.21
18:00
19.10.21
21:00
20.10.21
00:00

AI FOR COST MANAGEMENT

AI For Cost Optimization

- Instance Type Recommendation- recommends the best instance type based on workload patterns
- Cost-Aware Resource Scaling - AI scales resources while maintaining monetary constraints.





AI FOR RESOURCE MANAGEMENT

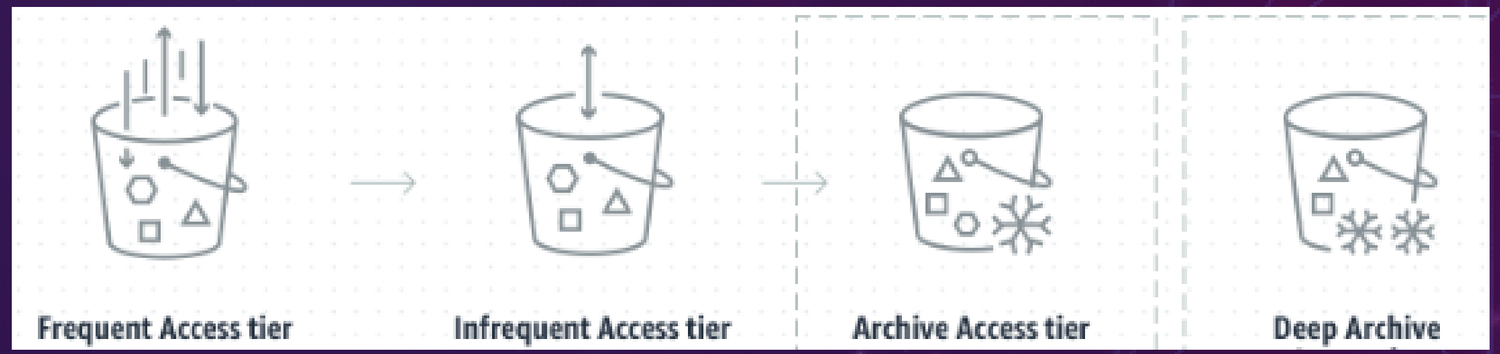
Storage and Data Management

- Intelligent Storage Tiering
- Data Access Prediction and Caching
- Data Compression and Deduplication
- Optimized Data Processing

AI FOR STORAGE AND DATA MANAGEMENT

Intelligent Storage Tiering

- Analyzing access patterns
  - Frequency of access
  - Volume of data
  - Session duration



- Important to
  - Utilize the full range of hardware.
  - Cost effective.

AI FOR RESOURCE MANAGEMENT

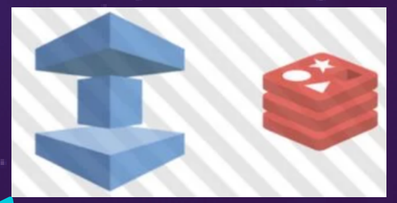
Storage and Data Management

- Intelligent Storage Tiering
- Data Access Prediction and Caching
- Data Compression and Deduplication
- Optimized Data Processing

AI FOR STORAGE AND DATA MANAGEMENT

Data Access Prediction and Caching

- Further improve data retrieval time for time-sensitive data



- ML features
  - The geographical region.
  - Application type
  - Access Frequency and Time-Based Access Patterns.

AI FOR RESOURCE MANAGEMENT

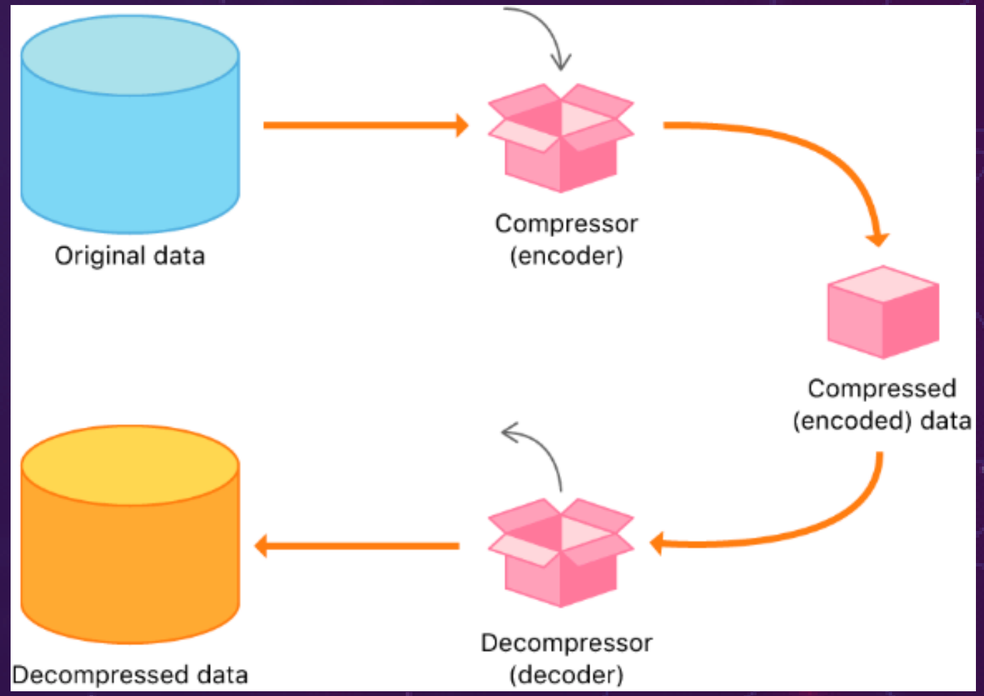
Storage and Data Management

- Intelligent Storage Tiering
- Data Access Prediction and Caching
- Data Compression and Deduplication
- Optimized Data Processing

AI FOR STORAGE AND DATA MANAGEMENT

Data Compression and Deduplication

- Most efficient format to maximize storage utilization.
- Store unique data only.
- Handle larger datasets without increasing storage requirements.



AI FOR RESOURCE MANAGEMENT

Storage and Data Management

- Intelligent Storage Tiering
- Data Access Prediction and Caching
- Data Compression and Deduplication
- Optimized Data Processing

AI FOR STORAGE AND DATA MANAGEMENT

Optimized Data Processing

- Select the best combination of services based on
  - Workload’s size
  - Urgency
  - Resource needs.

- Processing options
  - Serverless Execution.
  - Batch Processing.
  - Stream Processing.

AI FOR RESOURCE MANAGEMENT

AI for Security and Fault Tolerance

- AI for Predictive Maintenance
- Self-Healing Systems
- Proactive Security Resource Scaling

AI FOR SECURITY AND FAULT TOLERANCE

AI for Predictive Maintenance

- Monitor the health of cloud infrastructure in real-time.
- Identify optimal times for scheduled maintenance.
- Reduce unplanned interruptions.

PRESS RELEASE

.conf24: Splunk Report Shows Downtime Costs Global 2000 Companies $400B Annually

AI FOR RESOURCE MANAGEMENT

AI for Security and Fault Tolerance

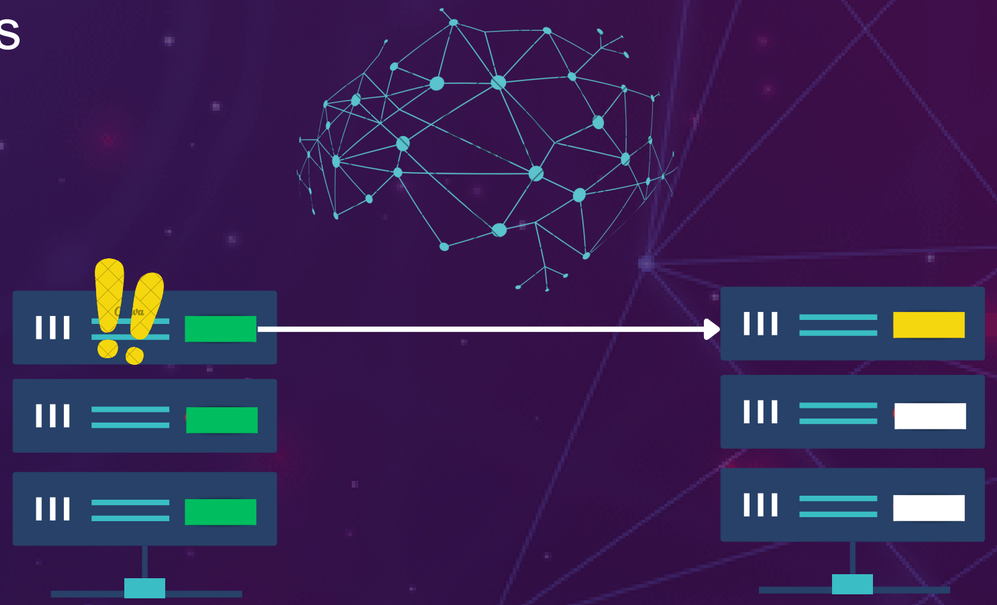
- AI for Predictive Maintenance
- Self-Healing Systems
- Proactive Security Resource Scaling

AI FOR SECURITY AND FAULT TOLERANCE

Self-Healing Systems

- Detection system for failures.

- Migrate workloads to healthier environments
- Reconfigure system resources.



AI FOR RESOURCE MANAGEMENT

AI for Security and Fault Tolerance

- AI for Predictive Maintenance
- Self-Healing Systems
- Proactive Security Resource Scaling

AI FOR SECURITY AND FAULT TOLERANCE

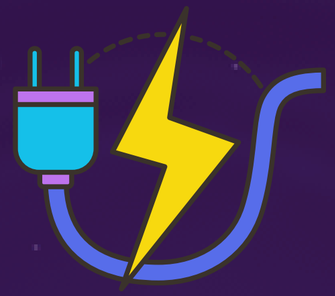
Proactive Security

Resource Scaling

- ML models can analyze real-time threat data.
- Predict potential security risks.
- Scale security measures.

AI FOR RESOURCE MANAGEMENT

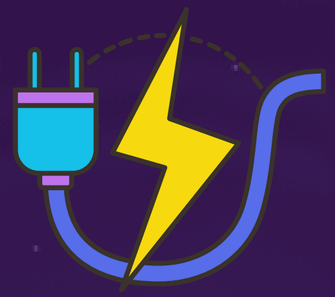
AI-Enabled Energy Efficiency



- Physical Machine Optimization
- Energy-Aware Workload Scheduling
- Cooling System Optimization

AI-ENABLED ENERGY EFFICIENCY

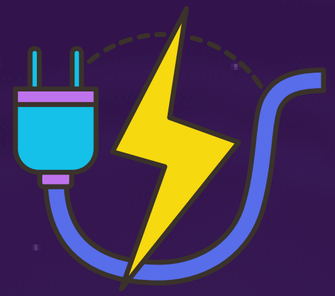
Physical Machine Optimization



- Predict near future demands.
  - Low demand >> consolidate work onto fewer servers.
  - Peak hours >> spread out to prevent resource contention or service degradation.

AI FOR RESOURCE MANAGEMENT

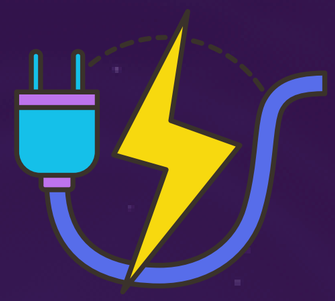
AI-Enabled Energy Efficiency



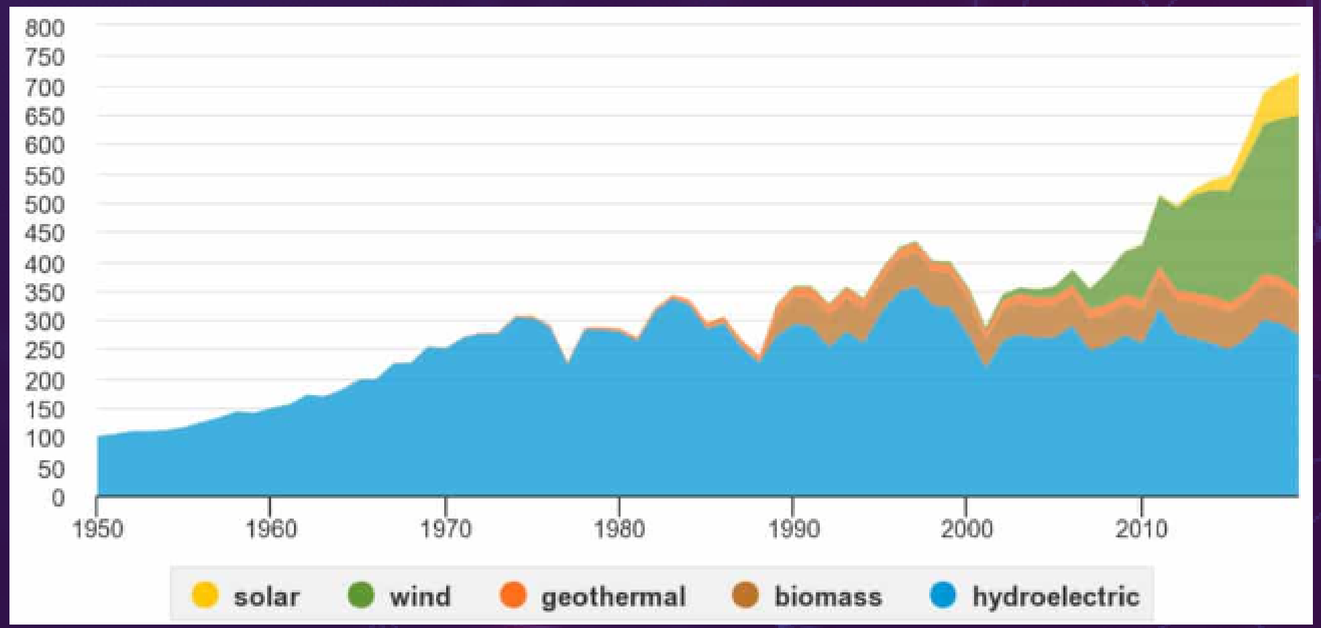
- Physical Machine Optimization
- Energy-Aware Workload Scheduling
- Cooling System Optimization

AI-ENABLED ENERGY EFFICIENCY

Energy-Aware Workload Scheduling

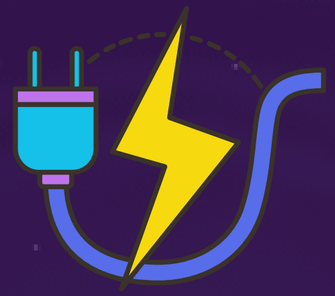


- Consider energy availability and cost when scheduling tasks.
- Prioritizing workloads when renewable energy is abundant or electricity rates are lower



AI FOR RESOURCE MANAGEMENT

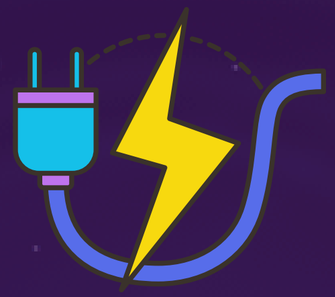
AI-Enabled Energy Efficiency



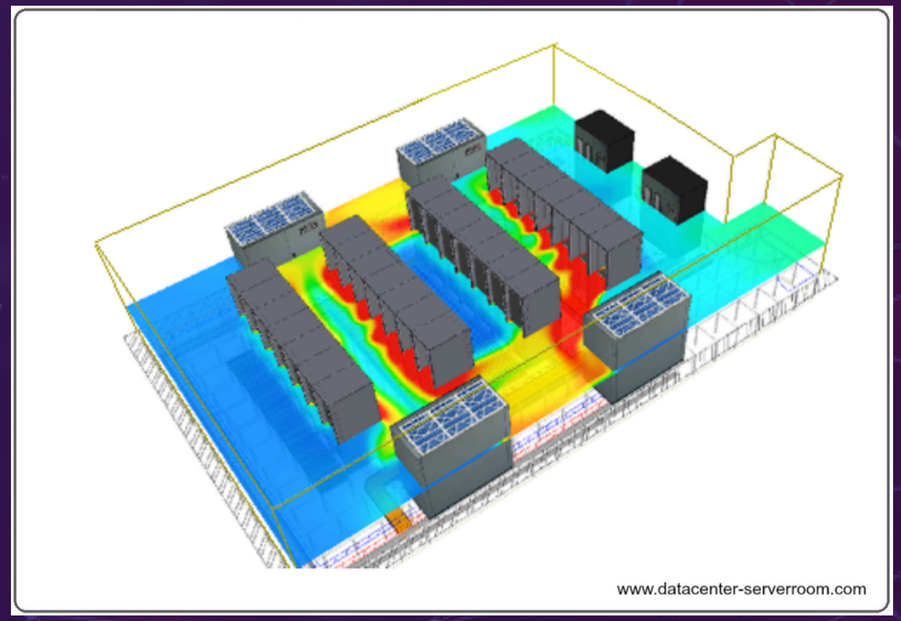
- Physical Machine Optimization
- Energy-Aware Workload Scheduling
- Cooling System Optimization

AI-ENABLED ENERGY EFFICIENCY

Cooling System Optimization



- Predict heat generation in data centers based on workload patterns.
- Dynamically adjust cooling intensity.



CONCLUSION

Artificial Intelligence can and should be leveraged for better resource management, as it not only enhances efficiency but also amplifies the core benefits that cloud computing already provides.

COST OPTIMIZATION

Instance Type Recommendation

Cost-Aware Resource Scaling

ENERGY EFFICIENCY

Dynamic Energy Consumption Balancing

Energy-Aware Workload Scheduling

STORAGE MANAGEMENT

Intelligent Storage Tiering

Data Compression and Deduplication

Data Access Prediction and Caching

SECURITY & FAULT TOLERANCE

Predictive Maintenance

Self-Healing Systems

Proactive Security Resource Scaling

REFERENCES

Graham, C., Upadhyay, S., Cheparthi, A., & Schumacher, R. (2023). Forecast: Public cloud services, worldwide, 2021-2027, 3Q23 update.

Mandal, A. (2024). Cloud resource management: An ultimate guide. Lucidity. https://www.lucidity.cloud/blog/cloud-resource-management

Yavorovych, D., Khoma, Y., Roubalik, Z., & Kerkhove, T. (2022, February 14). Introducing PredictKube - An AI-based predictive autoscaler for KEDA made by Dysnix. KEDA. https://keda.sh/blog/2022-02-09-predictkube-scaler

Chokkappagari, R. (2024) The role of ai and machine learning in cloud storage. Insights2Techinfo.

Sharma, D., Kumar, A., Tyagi, N., Chavan, S. (2023) Towards intelligent industrial systems: A comprehensive survey of sensor fusion techniques in iiot. Measurement: Sensors

Sathupadi, K. (2023) Ai-driven energy optimization in sdn-based cloud computing for balancing cost, energy efficiency, and network performance. International Journal of Applied Machine Learning and Computational Intelligence, 13(7):11–37.

THANK YOU!