Connect Beyond Cloud

Open AI Networking Korea 2022

클라우드 운영을 위한 인공지능 AIOps

윤호영 연구소장

22.11.10. (수)

OKESTRO

Your Future Cloud, Cloud Management Platform

OKESTRO

Open AI Networking Korea 2022

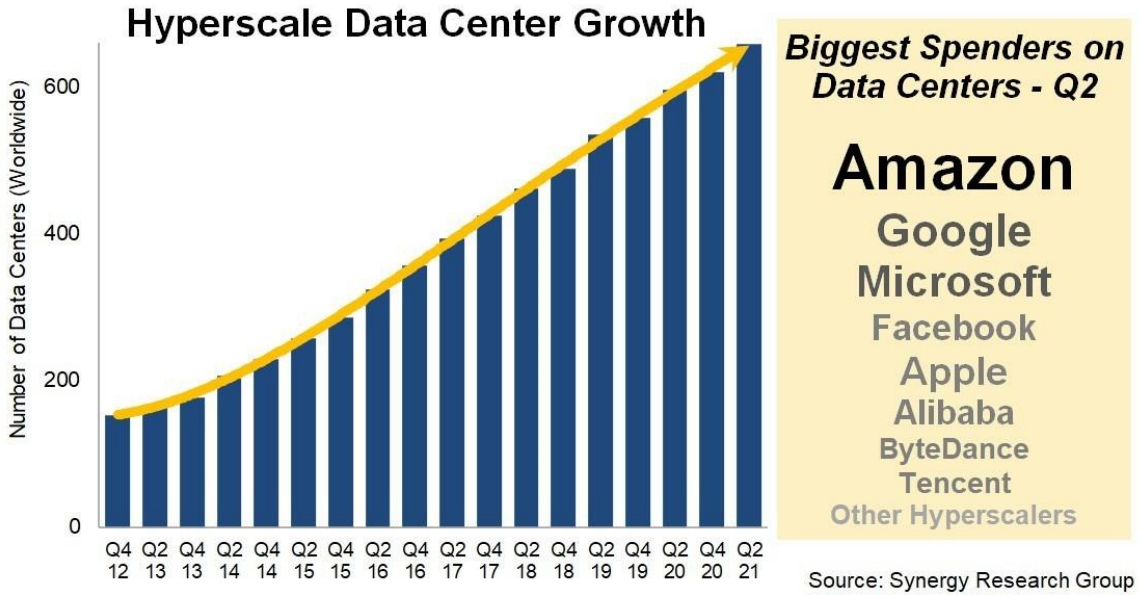
CONTENTS

1 데이터센터의 현황

● 클라우드 운영 환경을 제공하기 위한 클라우드 데이터센터의 증가

➢ 하이퍼스케일(Hyper Scale) 규모의 대형 데이터센터로 확장되고 있는 현황

- 데이터 사용량이 급증함에 따라, 전 세계 기업들이 데이터센터에 투자하는 비용도 함께 증가하고 있음
- Bigger is better의 운영공식을 갖기 때문에 전세계적으로 하이퍼스케일 데이터센터 수가 증가하고 있음
  (하이퍼스케일 데이터센터의 기준 : 최소 10만대 수준의 서버, 22,500㎡ 이상의 규모)
- 선두기업 : 아마존, 마이크로소프트, 알리바바, 구글, IBM 등



Source: Synergy Research Group

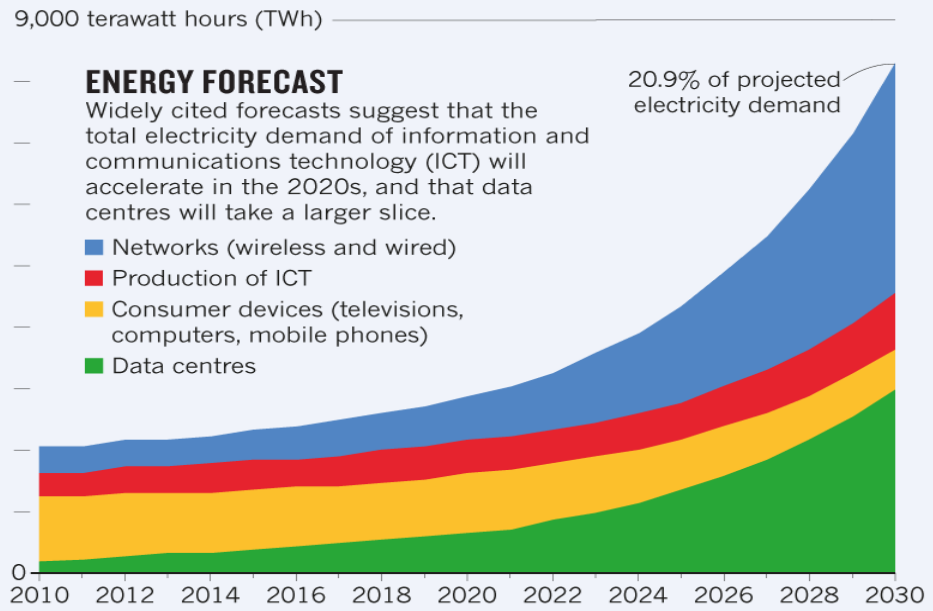
1 데이터센터의 현황

● 데이터센터의 문제점은 무엇인가?

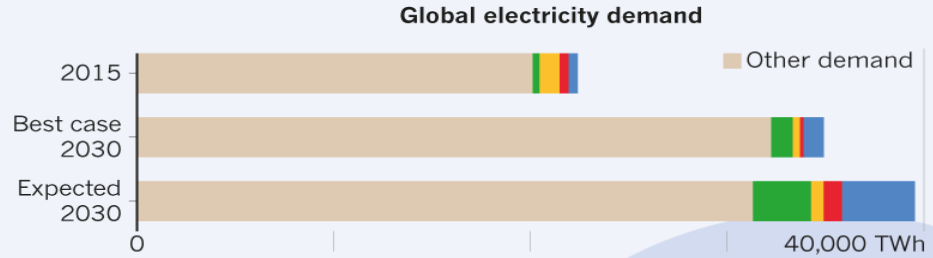
➢ 데이터센터는 전력 과소비 문제를 보이고 있음

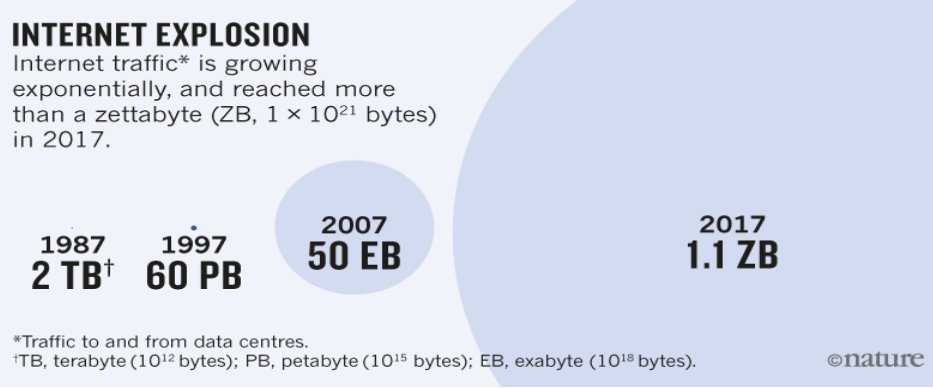
(≠ Clean Industry without Chimneys)

- 데이터센터에서는 서버, 네트워크 스토리지 등 IT 서비스 제공에 필요한 장비가 24시간 365일 운영되고 있음
- 다수의 인프라 장비들로부터 발생하는 발열을 줄이기 위해 공조, 냉각시설 등의 별도의 전력 사용이 발생함
- 전 세계적으로 매년 약 200TWh의 전력을 사용하는 전력 사용량이 가장 높은 단일 건물로, OPEX가 CAPEX보다 높음
- 전 세계 전력 사용량의 약 2%를 차지하며, 일부 국가의 총 전력사용량보다 많은 양임
- ICT 기술 관련 전력 수요는 2020년대에 가속화될 것이며, 그 중 데이터센터가 큰 비중을 차지할 것으로 예측 - nature
  - 데이터 센터의 전력 사용량이 2030 년까지 약 15 배 증가하여, 전 세계 전력 사용량의 8 %까지 증가할 것으로 예상



The chart above is an 'expected case' projection from Anders Andrae, a specialist in sustainable ICT. In his 'best case' scenario, ICT grows to only 8% of total electricity demand by 2030, rather than to 21%.



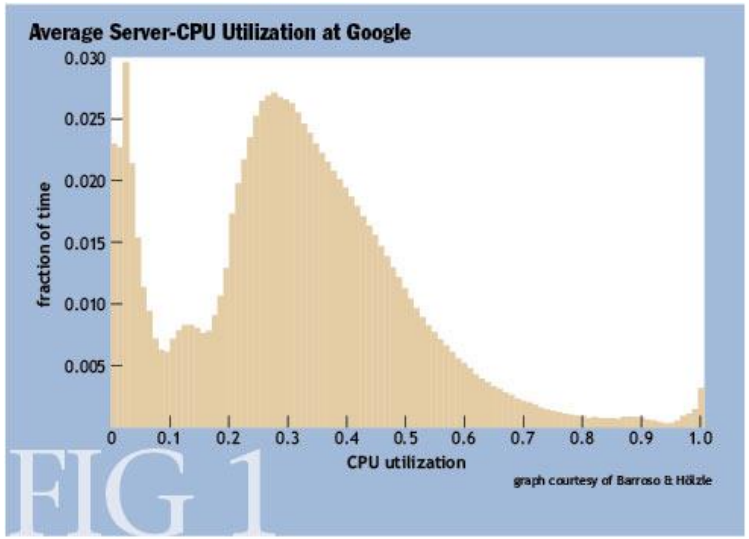


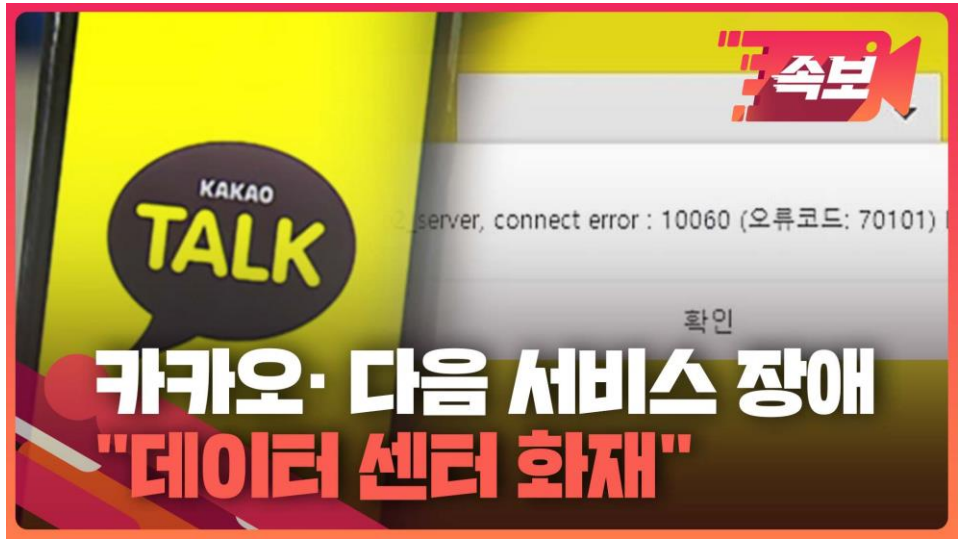
1 데이터센터의 현황

● 데이터센터는 비효율적인 운영으로 인한 전력 낭비가 심각함

➢ 보수적인 운영방안 고수로 인한 데이터센터의 비효율적인 운영현황

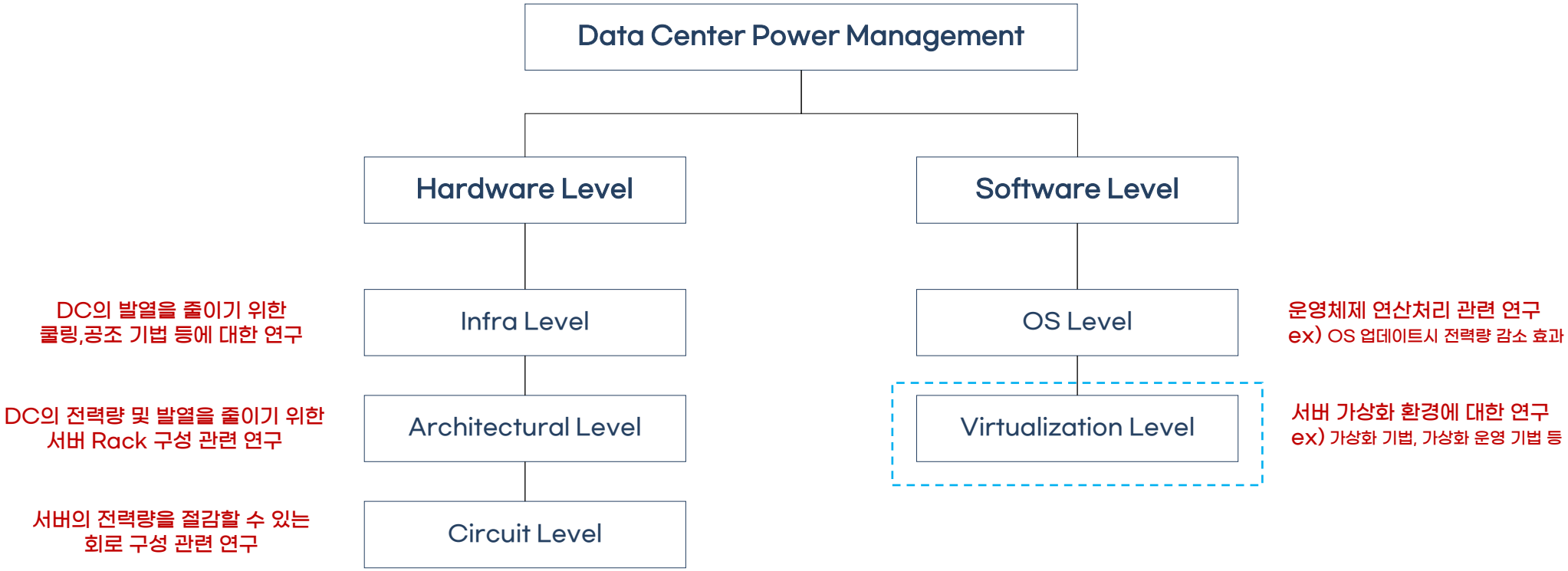
- 데이터센터에서 가동되는 서버의 평균 CPU 사용량은 10~15%에 불과함(Google Datacenter 기준)
- 국내 데이터센터의 전력효율지수(PUE; Power Usage Effectiveness)는 민간 2.66, 공공지자체 3.13 (해외 평균지수 1.64)
- 데이터센터가 가동이 중단(downtime)됐을 때, 피해비용은 1분당 약 $9,000로 추정됨(Gartner)
  → 국방, 항공 등의 Mission Critical 시스템을 운영하는 데이터센터의 가동 중단은 국가재난급의 사태로 번질 수 있음
- 데이터센터는 계약한 고객과의 SLA 준수를 위해 안정성을 최우선시하고, 전력 효율성을 고려하지 않는 경향이 있음





2 Literature Review

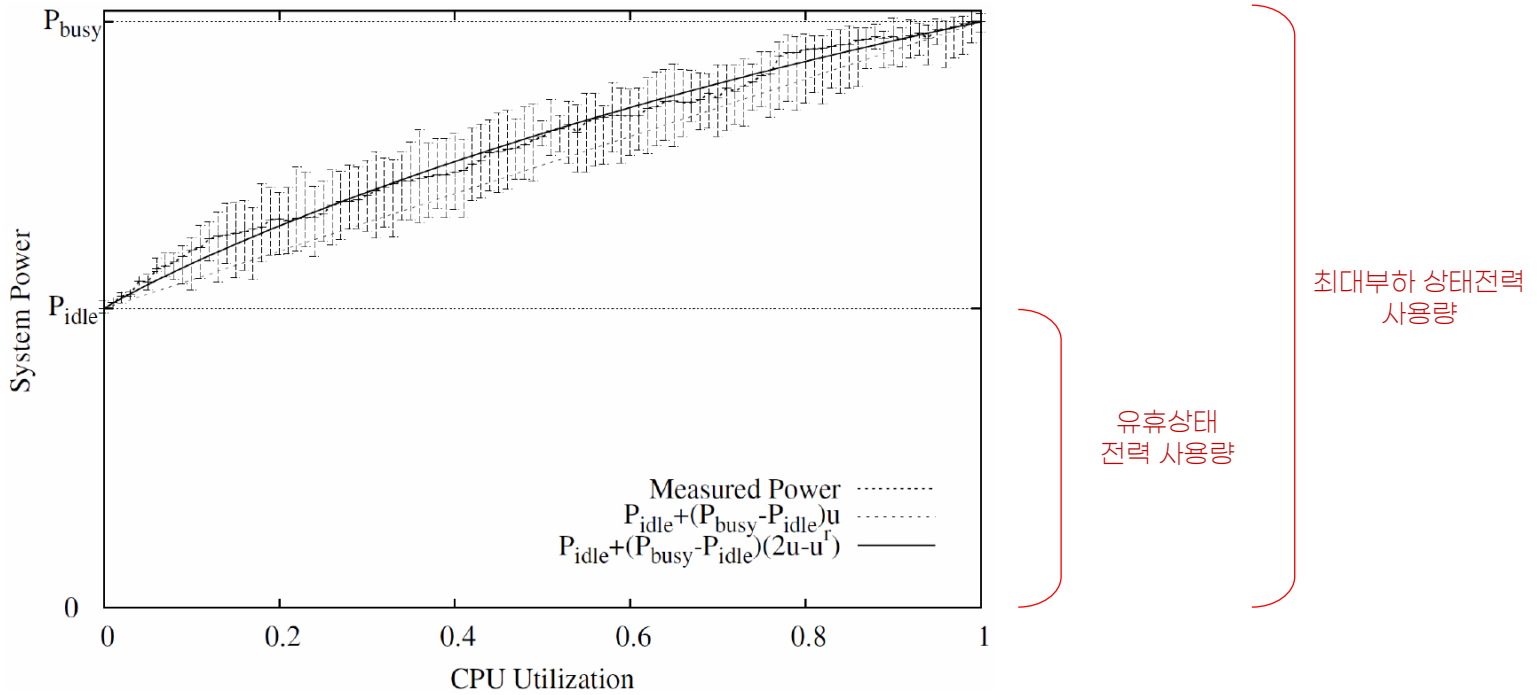
● 데이터센터 전력 관리에 대한 연구



Key Point	데이터센터 전력 관련 연구는 크게 쿨링 시스템, 설비 구조, 회로도 개선 등 H/W 관련 연구와 운영체제(OS)와 가상화(Virtualization) 기술 등의 S/W 관련 연구로 나눠지며, H/W 관련 기법은 데이터센터 설계 당시에 고려하거나 추가 설비가 투입되어야 하는 반면, S/W는 이미 구축된 환경에도 유연하게 적용 가능한 특징이 존재함

2 Literature Review

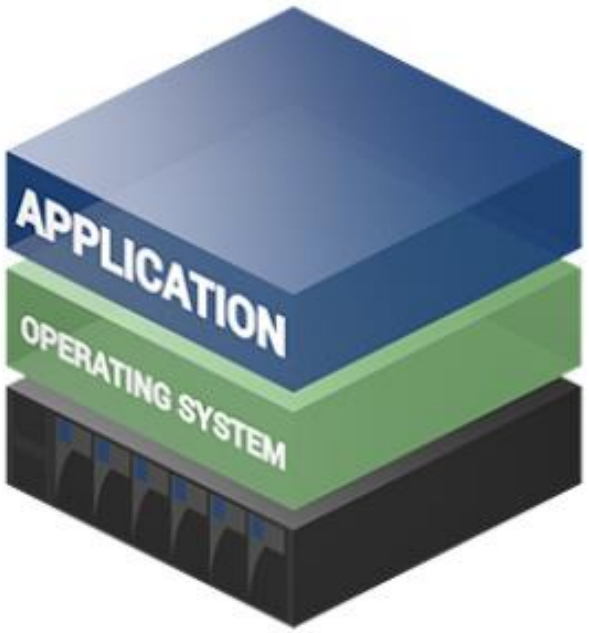
- **데이터센터의 전력량을 줄이는 가장 효과적인 기법 ➔ 서버의 Shutdown**
  - 가동 중인 서버를 **shutdown**시키는 것이 데이터센터의 전력 절감 기법 중 가장 효과적인 기법임(Beloglazov, 2013)
    - 데이터센터의 H/W, S/W, OS Level의 전력 절감 관련 연구를 대상으로 조사한 결과



Key Point	서버는 아무 작업을 하지 않은채로 켜져만 있어도, 기본적으로 많은 전력을 소모하기 때문에 전력량을 줄이기 위해서는 shutdown시키는 것이 가장 효과적임. 위 차트는 CPU의 점유율이 0에 가까운 상태(아무 작업이 이루어지지 않은 상태)여도, P_{idle} 만큼의 전력을 소모하는 것을 나타냄

Literature Review

- **가동 중인 서버를 어떻게 shutdown 시킬 수 있는가?**
  - **가상화(Virtualization)** : 하나의 서버를 여러 개의 가상 환경(Virtual Machine, 이하 VM)으로 나누어 사용하는 기술



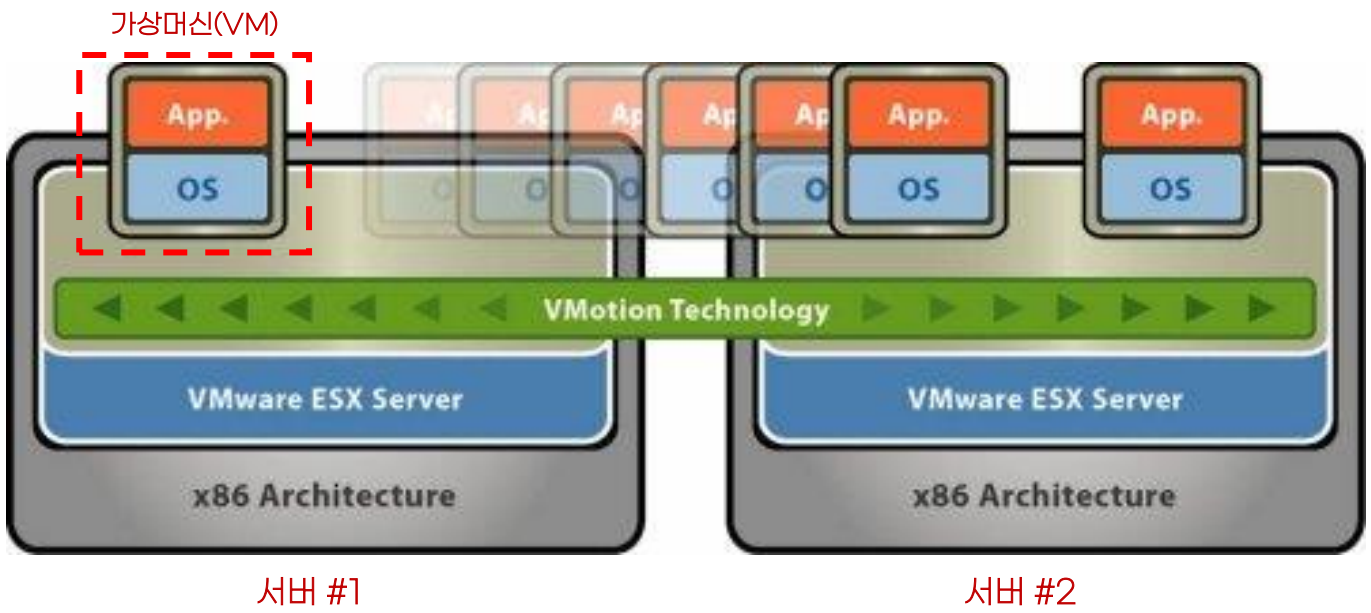
[Traditional Architecture]

[Virtualization Architecture]

Key Point	서버는 일반 데스크탑 보다 고사양의 장비로, 효율적으로 사용하기 위해 가상화(Virtualization) 기술을 적용하여 운영함 (좌) 1대의 데스크탑에 1개의 OS를 설치하여 사용하는 일반적인 방식 (우) 1대의 서버를 6개의 가상머신(VM)으로 나누어 사용하는 가상화 방식 - 서버의 자원(CPU, Memory 등)을 가상머신들이 나눠서 사용할 수 있음

Literature Review

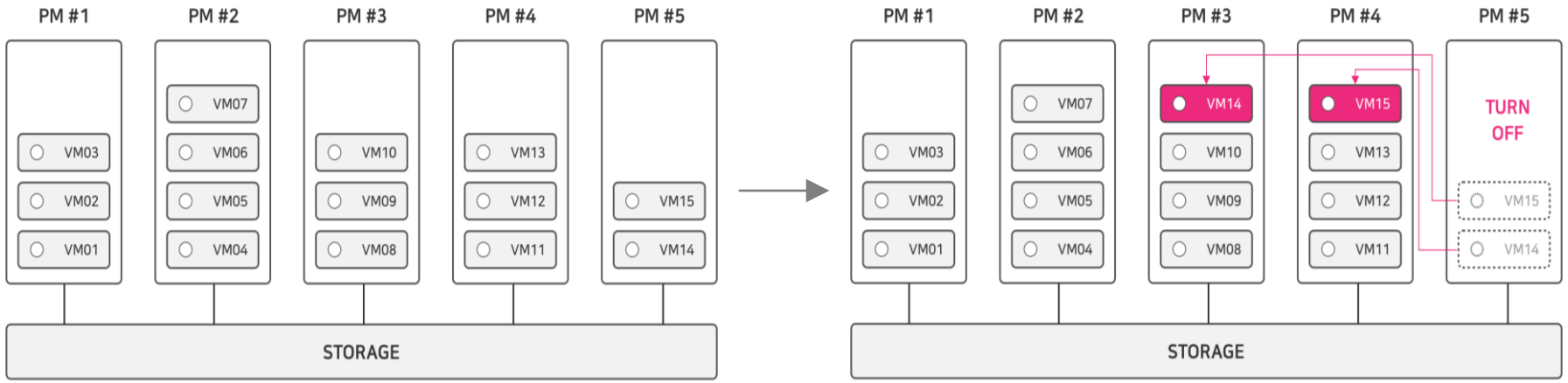
- **가상화 기술은 다른 서버로 VM을 옮길 수 있는 기능을 지원함**
  - **라이브 마이그레이션(Live Migration) :** 특정 서버에 위치한 VM을 다른 서버로 이동시키는 기술로 마이그레이션 도중에도 VM에서 실행중인 어플리케이션과 클라이언트의 접속이 끊기지 않고 유지되는 특징이 있음
  - 하지만, VM Migration의 행위 자체가 **과부하(overload)를 유발**하고, **미세한 down time이 발생**하는 단점이 존재하기 때문에 필요한 경우가 아니면 **지양해야 하는 작업**임



[서버 #1에 할당되어 있는 VM을 서버 #2로 migration 하는 예시]

2 Literature Review

- **VM Migration 기능을 활용하여, 서버를 유휴전환 하는 기법**



➢ **VM Placement** : Live Migration 기능을 활용하여, VM 을 동적으로 마이그레이션하여, 특정 목적을 달성하는 기법

- VM이 하나도 없는 서버(Physical Machine, 이하 PM)는 shutdown 시킬 수 있음 → 전력 사용량 절감을 유도
- VM Placement 는 특정 PM에 위치한 특정 VM을 다른 어떤 특정 PM으로 migration 시킬지를 결정하는 문제로 구성되어 있음

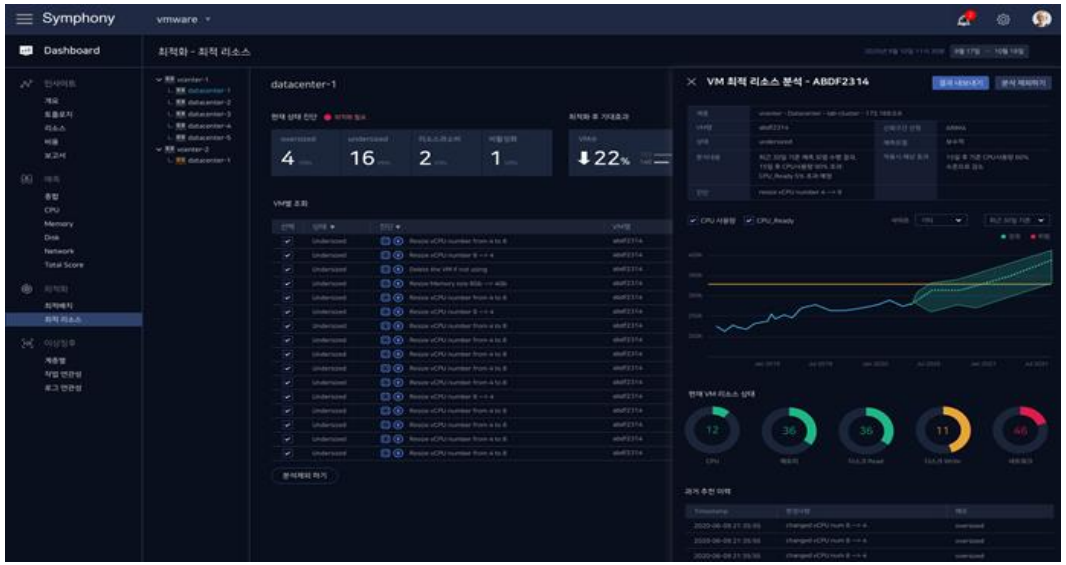
Key Point	VM Placement는 VM을 적재적소에 다른 PM으로 마이그레이션하여, 가용 PM을 줄여 전력 사용량의 절감을 유도하고, PM의 utilization을 높이는 것을 목적으로 함 - Source PM Selection, VM Selection, Destination PM Selection를 결정하는 문제로 구성되어 있음

2 Literature Review

● 예측 분석(Predictive Analytics)을 통한 클라우드 자원의 최적 배치안 제시

최적 운영을 위한 상황별 배치안 선택

1. 안정성과 효율성을 종합적으로 고려한 자원 배치안 제시
2. 점수화된 워크로드 안정성 지표를 통해 배치안 선택 근거 확보
3. 추천 배치안 별 전력 소비량 등 비용 분석 기능 제공으로 비용 절감 측면에서 배치안 고려 가능

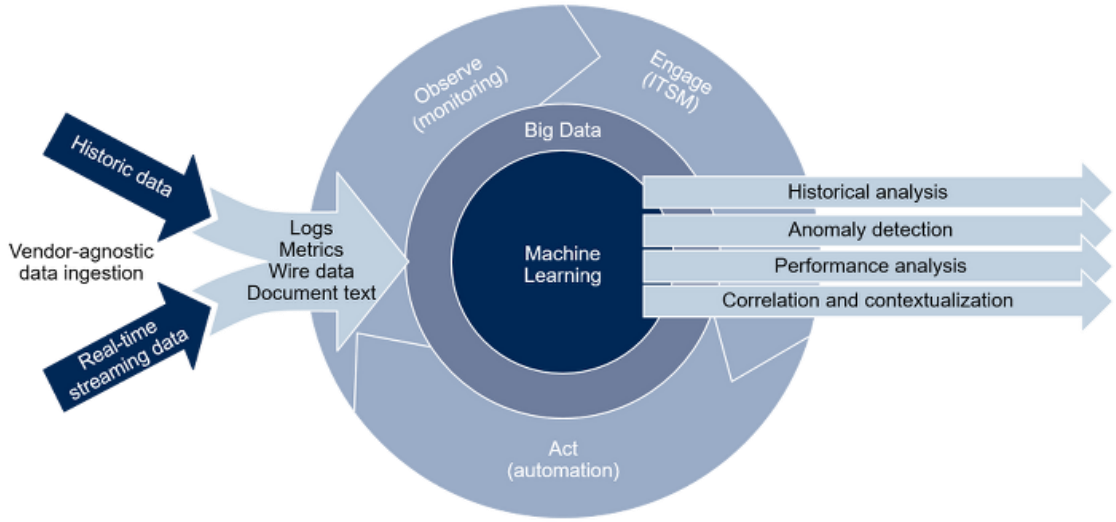


기대효과 확인 및 이력관리

1. 개별 VM별 상세정보 확인을 통하여 워크로드 예측 데이터 및 용량 산정 기준 시각화 제시
2. 용량 산정 기대효과를 확인하여 효율성 증가 정도 확인 가능
3. 인프라 변경사항에 대한 이력을 관리하여 종합적인 판단 가능

3 AIOps: AI for IT Operations

"지속가능한 IT 운영을 위한 AIOps 플랫폼"



*AIOps Platform Enabling Continuous ITOM | 출처 : Gartner

dynatrace

DATADOG

" 오늘날 운영/관제 대상 시스템이 다양해지고 메트릭, 로그 등 운영 데이터의 양도 급격히 늘어남에 따라 A.I.를 도입하여 안정적 서비스 보장과 운영 요원의 피로도 감소, 운영 비용 절감 등을 꾀하는 것을 목적으로 하고 있음 "

"AIOps는 10번째로 유력한 기술 및 트렌드가 될 것으로 전망"

IDC FutureScape 2019
Race To Reinvent For Multiplied Innovation

IDC
ANALYZE THE FUTURE

2019 국내 ICT시장 10대 전망

1. 디지털 디터미네이션(Determination)
2. 데이터 수익화(Monetization)
3. 디지털 KPI
4. 디지털 트윈
5. 애자일 연결성(Connectivity)
6. 블록체인 기반의 DX 플랫폼
7. 엣지(Edge) 영역으로의 확장
8. 앱데브(AppDev) 혁명
9. 새로운 UI로서의 AI
10. AI기반의 IT 운영(Operations)

For more information, visit www.idc.com/futurescapes2019korea
@IDCKorea IDC KOREA

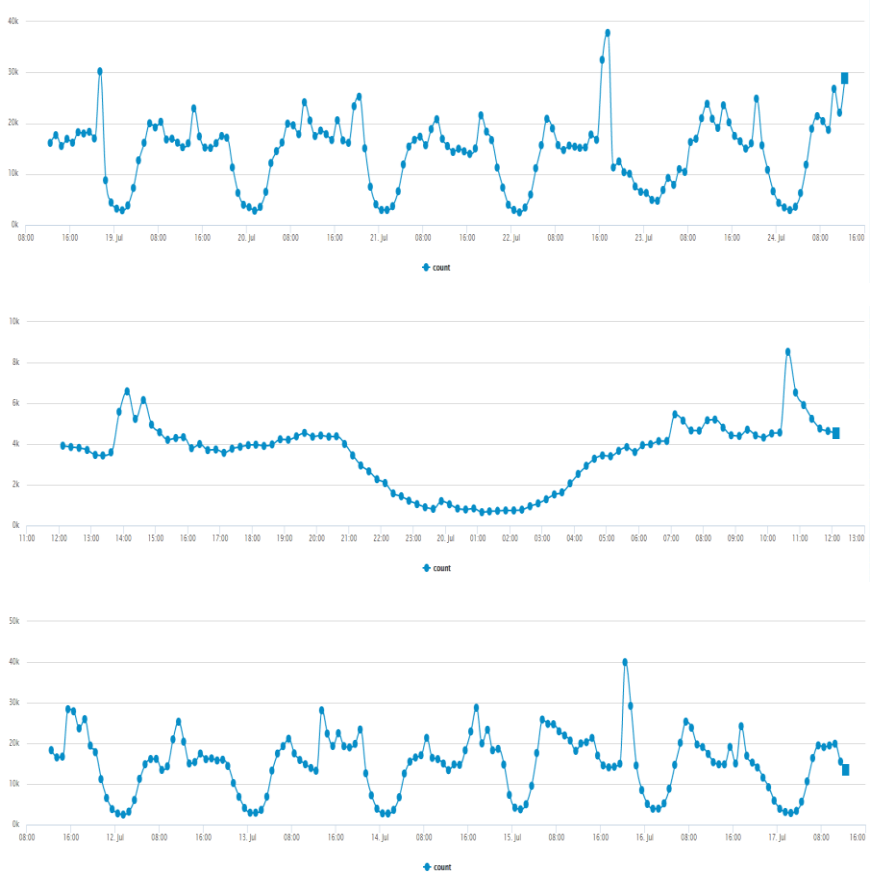
#IDCFutureScapes

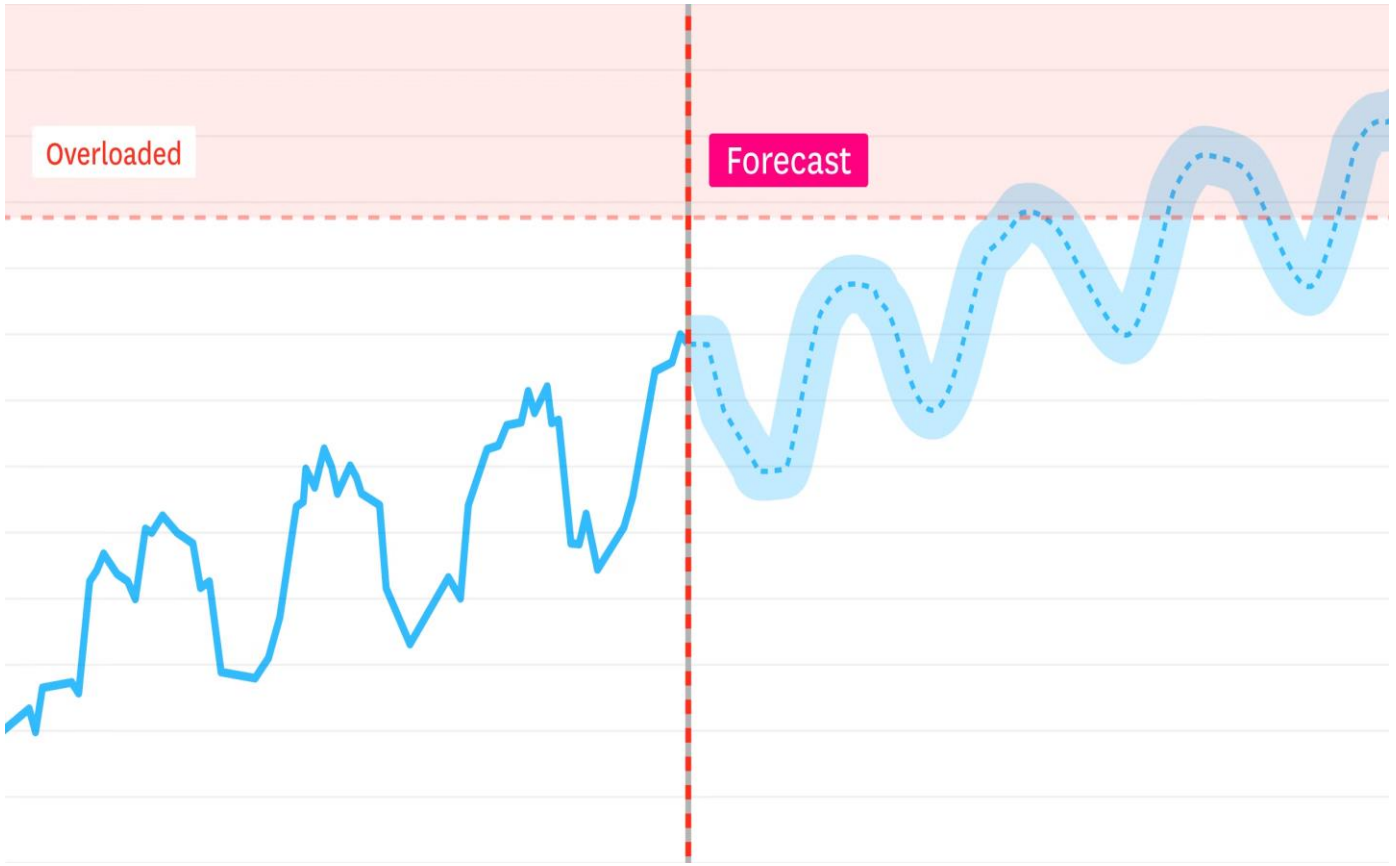
* 2019 국내 ICT 시장 10대 전망 | 출처 : 한국IDC

" AIOps를 통해 IT 지출을 축소하고, 기업의 IT 민첩성을 개선하며, 혁신을 가속화 할 수 밖에 없게 되면서, 60%의 CIO가 2021년까지 IT운영, 툴, 프로세스에 있어 데이터 및 A.I.를 공격적으로 적용하게 될 것 "

3 AIOps: AI for IT Operations

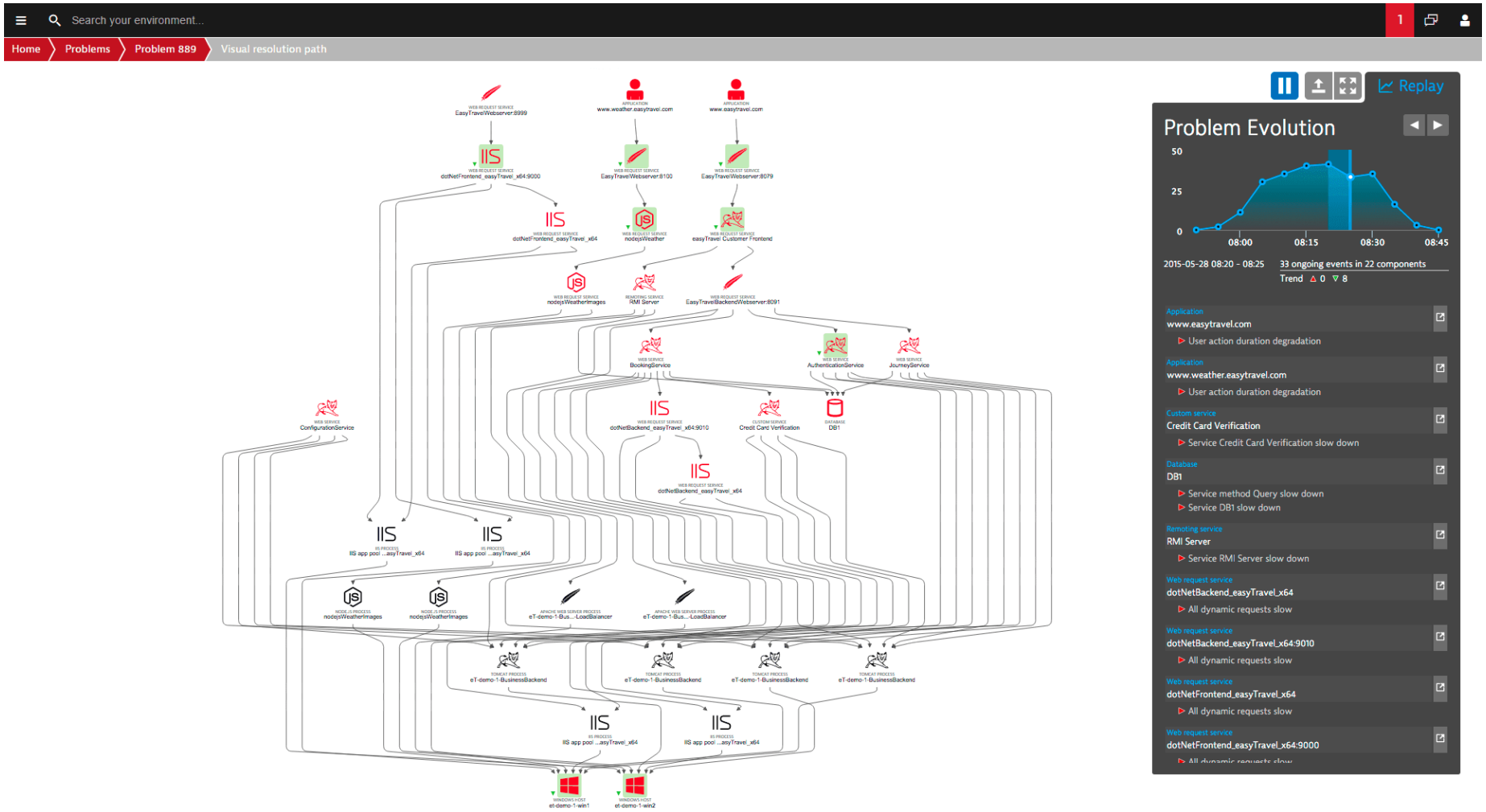
- **Prediction-based Anomaly Detection**





3 AIOps: AI for IT Operations

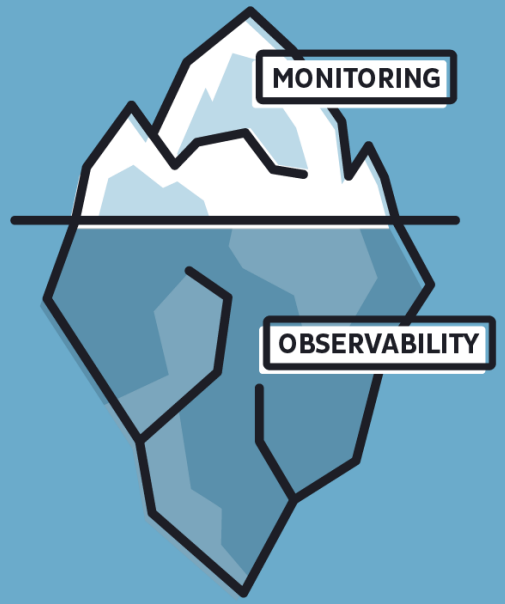
- **근본 원인 분석(RCA: Root Cause Analysis)**



3 AIOps: AI for IT Operations

- **Beyond Monitoring: The Rise of Observability**

Relationship Between Observability and Monitoring

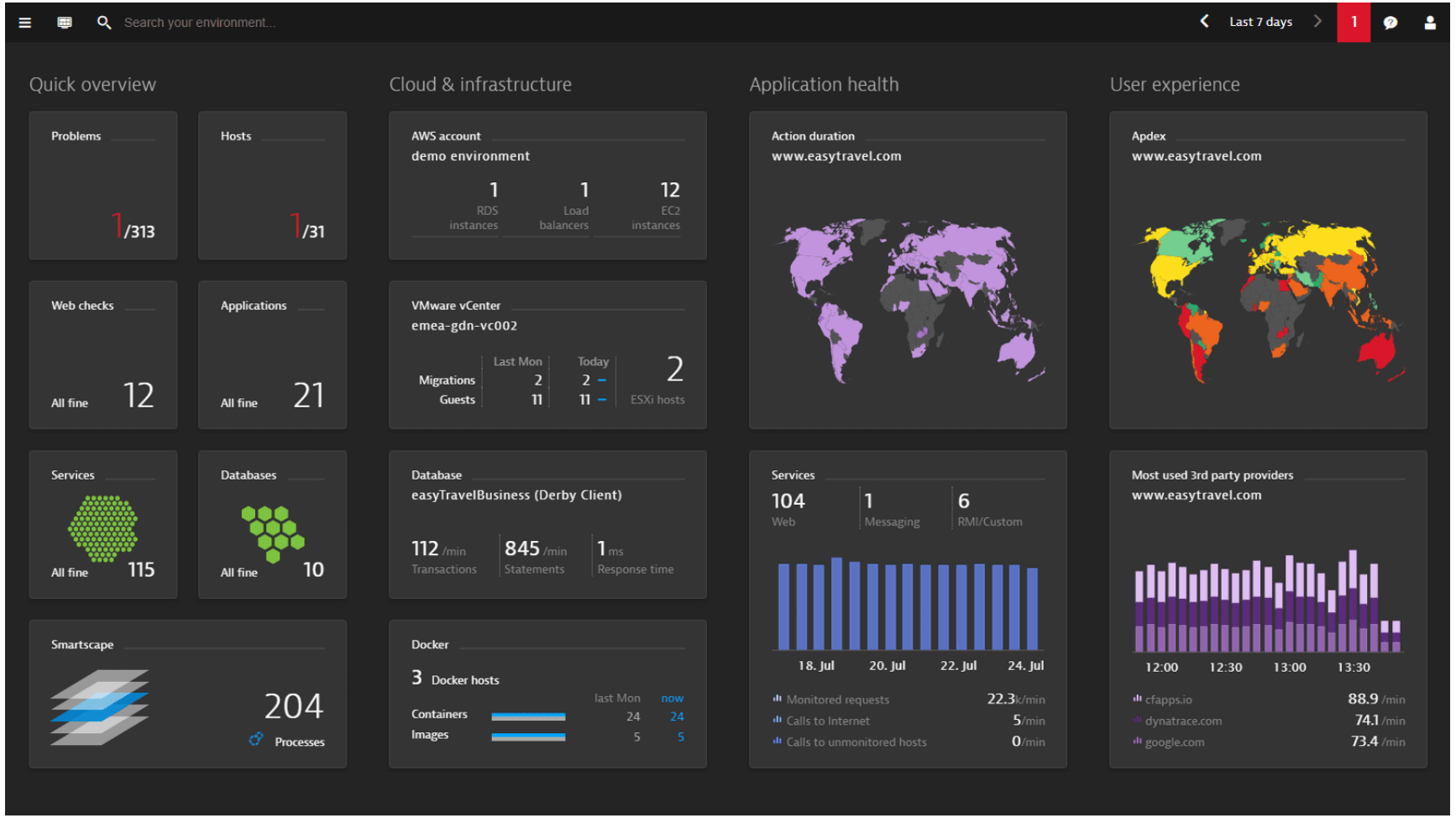


While observability and monitoring both rely on similar data - metrics, traces and logs - monitoring is reactive while observability allows proactive prevention.

Observability	Monitoring
A system has to be designed to be observed	Almost anything can be monitored
What's my system doing?	Is my system working?
Passive (Push/Publish)	Active (Pull/Collect)
Includes traces	Includes metrics, events, logs
Helps isolate issues at scale	Challenging at scale
Relies heavily on sampling	Can provide raw data
Reduces duration of outages	Provides rapid response to outages
Generate metrics	Collect metrics
Why my system failed	What's the state of my system

3 AIOps: AI for IT Operations

- **Beyond Monitoring: The Rise of Observability**



4 Conclusion

- **완전 자동화(Self-optimizing, autonomic) 이전의 최적화(Optimizing)**

Data center management maturity model

LEVEL	DESCRIPTION	OPERATING EFFICIENCY	SOFTWARE
Level 5: Self-optimizing, autonomic	AI-driven integrated management software adjusts data center behavior and makes best use of resources according to goals, rules and service requirements throughout its lifecycle.	HIGH	AI-driven, integrated DCIM with automation
Level 4: Optimizing	Physical and virtual IT and data center subsystems integrated; models used for prediction, service management and multiple views, optimizing in near real time. AI is applied to DCIM-based data lakes for advanced analytics.	MEDIUM	AI-driven, integrated DCIM
Level 3: Proactive	Physical data center equipment characteristics, location and operational status is tracked. Energy and environmental data is used to reduce risks and waste.	MEDIUM	Integrated DCIM
Level 2: Reactive	Software installed to monitor environmentals and equipment power use. Able to adjust basic controls (e.g., cooling) to demand.	LOW	DCIM monitoring
Level 1: Basic	No integration of infrastructure data. Basic monitoring supplied with equipment. Relies on BMS data. Simple alarming, error messaging.	LOW	Ad hoc

by Rhonda Ascierto, Vice President, Research, Uptime Institute