



매니코어 기반 초고성능 스케일러블 OS 기초 연구

@ 2019-2차 매니코어 포럼

2019. 8. 14.
정성인, ETRI

SW기초연구과제 추진 (미래부 2014)

SW기초연구과제 (미래부 SW기초연구센터 추진방안에서)	기존의 연구과제
<ul style="list-style-type: none"> • (정의) TRL 2 (개념정립) ~ TRL 5 (요소기술로 시작품제작 및 검증) 단계로 미래 SW 기술 확보와 First-Mover 동력이 될 SW 기술 연구 • (연구영역) 단기 시장을 목표로 하지 않고, 기술적 도전 또는 신개념 창출을 목적으로 하는 이론과 실험적 연구 추진 <ul style="list-style-type: none"> - 기초이론과 알고리즘 연구 - 예시, 진입장벽이 높은 기술 (DB, 인공지능 등), 새로운 알고리즘, SW융합연구, 공개SW • (추진 방법-1) 기술 도전을 위해서 다양한 방법론을 제시하고, 동일한 주제를 수행하는 기관 구성 • (추진 방법-2) 연구분야별 국내외 전문가 영입 및 자체 연구인력으로 구성 • (기대효과) SW 기초 연구 ⇒ 핵심기술개발 ⇒ 창업 	<ul style="list-style-type: none"> • TRL3/4 ~ TRL 7/8 (상용 시제품) • 활용시나리오/요구사항정의-설계-구현-통합-시험 • 사업화

❖ 2개 SW 센터 사업 (2014~), 29개의 SW 스타랩 사업으로 추진 (2015~)

목차

- 매니코어 포럼 목적
- 과제 개요
- 매니코어 등 고성능 하드웨어 현황
- 문제점
- 연구개발 로드맵과 결과물
- 활용 방안 (유스 케이스)

매니코어 포럼의 목적

- 목적 1 : 연구개발 방향 모색 (1단계, 2014~2017)
 - 활용 방안 모색 (2단계, ~2021)
- 목적 2 : 파트너 모색
 - 유스 케이스 (use case) 구축할 파트너
 - 활용 측면 자문 파트너
 - 공개한 요소기술을 활용할 파트너
 - 오픈 프로젝트를 운영할 파트너 등

과제 개요

- 미래부 SW기초연구과제 (2014 ~)
 - SW기초연구센터 과제 (2개), 스타랩과제 (29개), 차세대과제 (10여개)
- 매니코어 기반 초고성능 스케일러블 OS 기초연구 (차세대OS기초연구센터)
 - 2014-2021, 공개SW과제
 - ETRI, 국민대, 건국대, 경성대, 부산대, 서강대, 서울대, 성균관대, 연세대, 전북대, 조지아공대, 버지니아공대, 에프에이리눅스, 락플레이스, 공개SW협회
 - 목표 : 코어 수 증가에 따른 OS 성능 증가 (스케일러빌리티 제공)
 - 대상 기술 : 리눅스 (모노리틱구조), 새로운 구조의 운영체제 (멀티커널 구조), 병렬화 환경, 도구
 - 오픈소스 활동 방식
 - 리눅스 커뮤니티에 피드백
 - 개발 후 공개
 - 오픈 프로젝트
 - <https://github.com/oslab-swrc>
 - <http://manycoreos.synology.me/mediawiki/index.php?title=ManycoreOS>

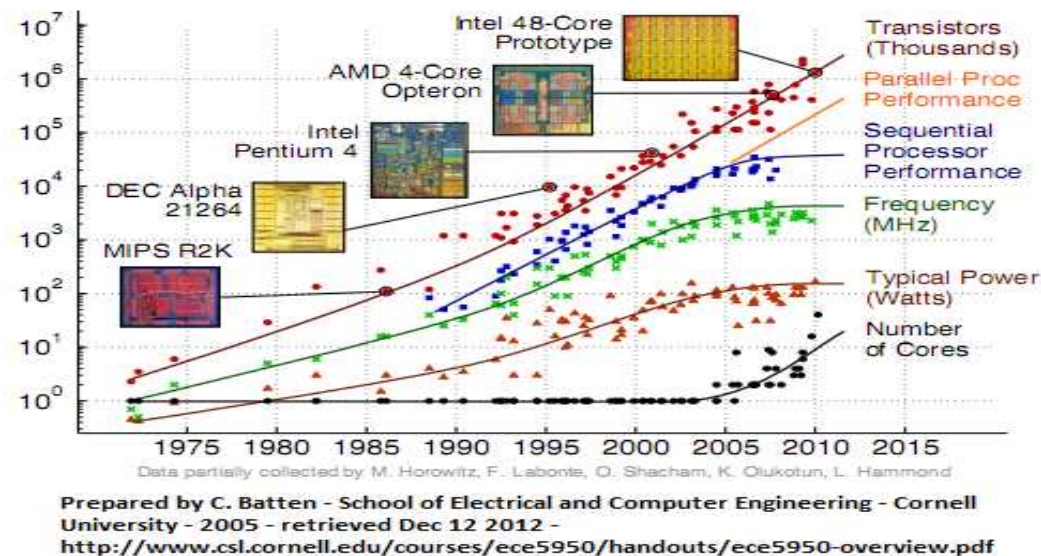
멀티코어 → 매니코어

- 배경

- 2004년에 프로세서 주파수 증가의 한계 봉착

- Intel, it's SW developers turn

- SW 개발자가 매니코어 환경을 고려하여 SW 개발 필요



매니코어 시스템



IBM 120 cores, 192 cores



Intel NVDIMM server 96 cores



ARM ThunderX 96 cores

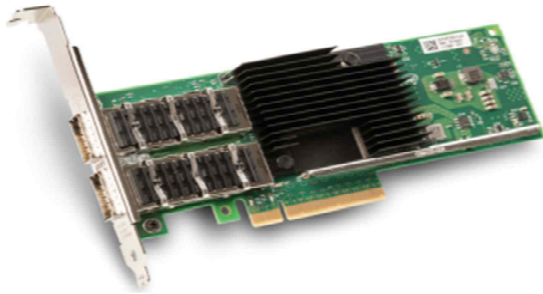


Intel Xeon Phi 7290F (KNL)
- 288 cores (72 cores * 4 thread)

고성능 스토리지

Device	Size	read IOPS	read xput (MB/s)	write IOPS	write xput (MB/s)
SATA HDD (HGST He10)	10TB	626	249	3,748	249
SATA SSD (850 Evo)	2TB	90,000	493	90,000	468
NVMe SSD (960 Pro)	1TB	440,000	3,500	360,000	2,100
NVMe 3D XPoint (Optane 900p)	480GB	550,000	2,500	500,000	2,000

고성능 NIC



Intel 40 GbE



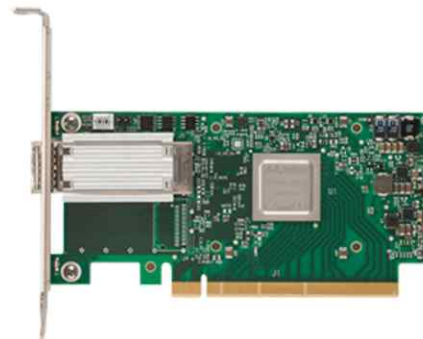
Mellanox 40 GbE



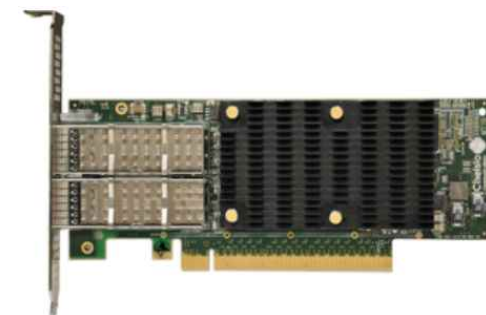
Chelsio 40 GbE



Intel 100 GbE



Mellanox 100

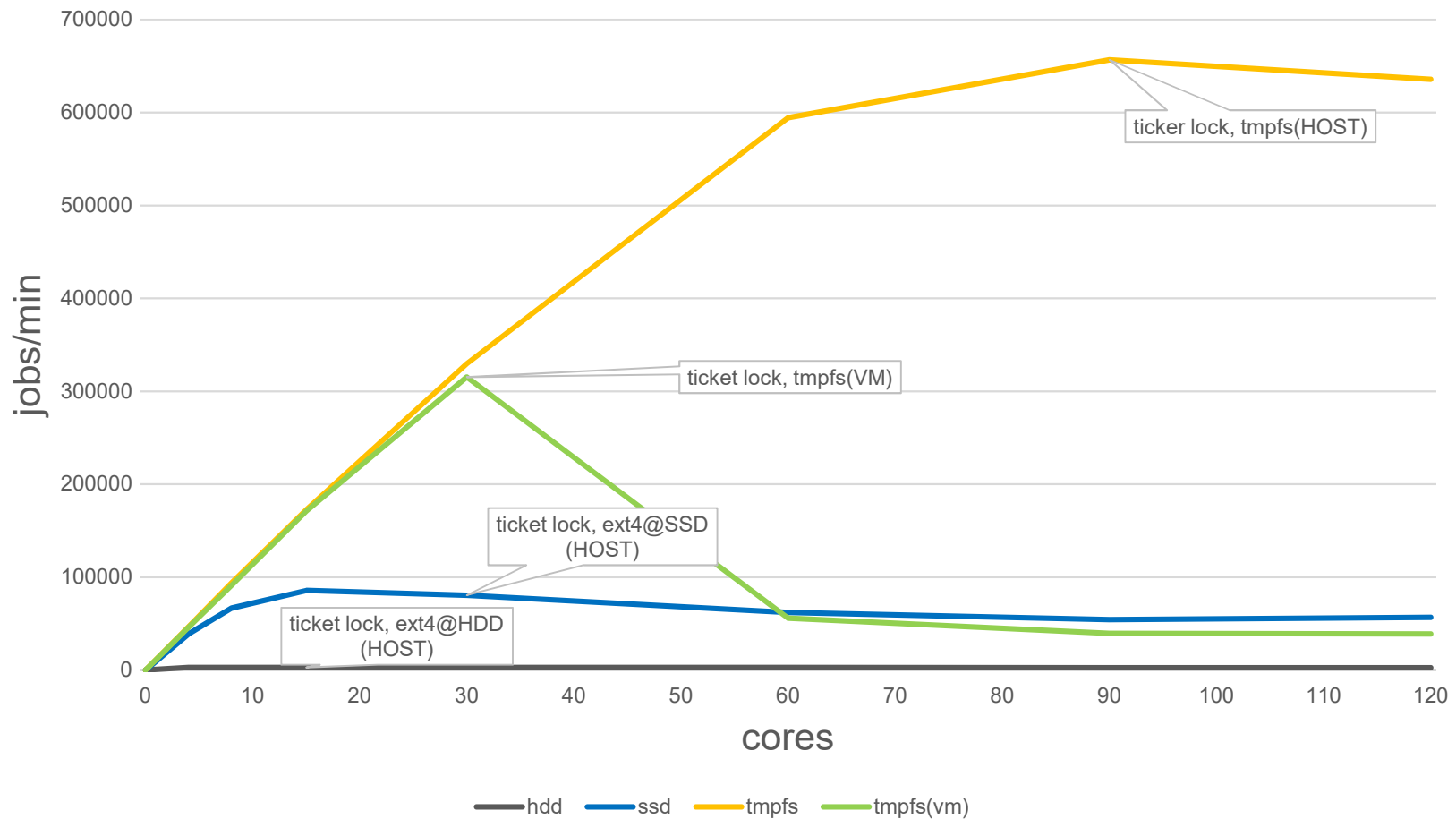


Chelsio 100 GbE

가진 소프트웨어의 성능을 어떻게 측정할까?

성능 그래프 (1)

- AIM7(OS 성능시험 도구) Multiuser 워크로드(계산:IO=50:50)로 성능 실험



성능 그래프 (2)

- MIT paper (2010), An Analysis of Linux Scalability to Many Cores

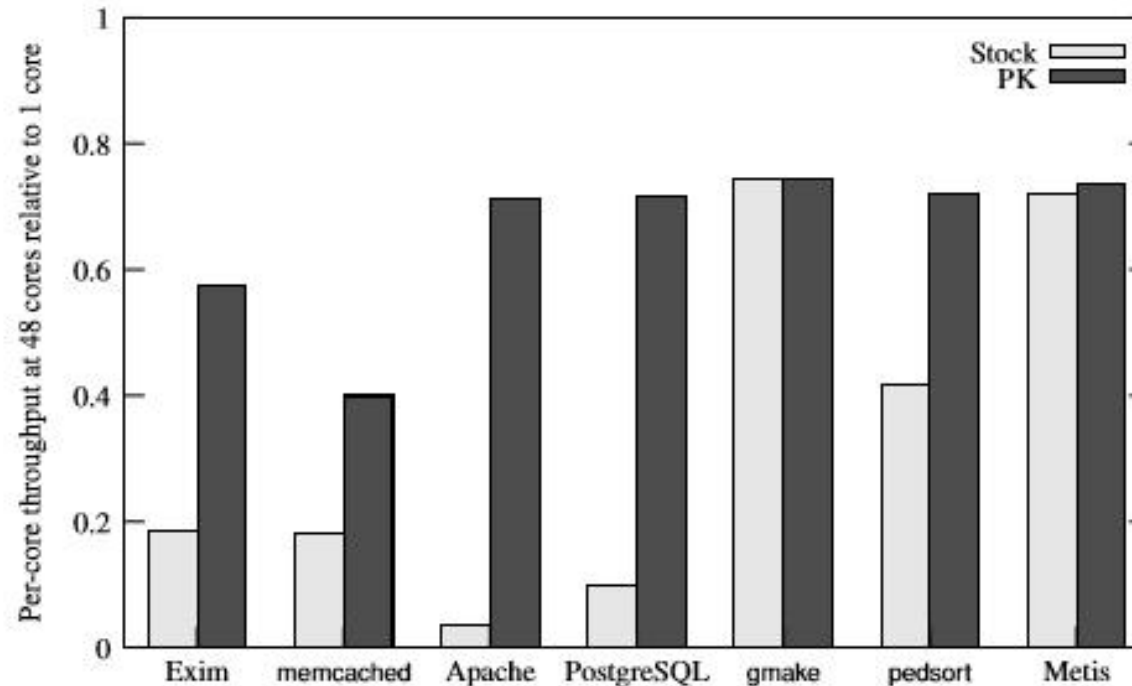
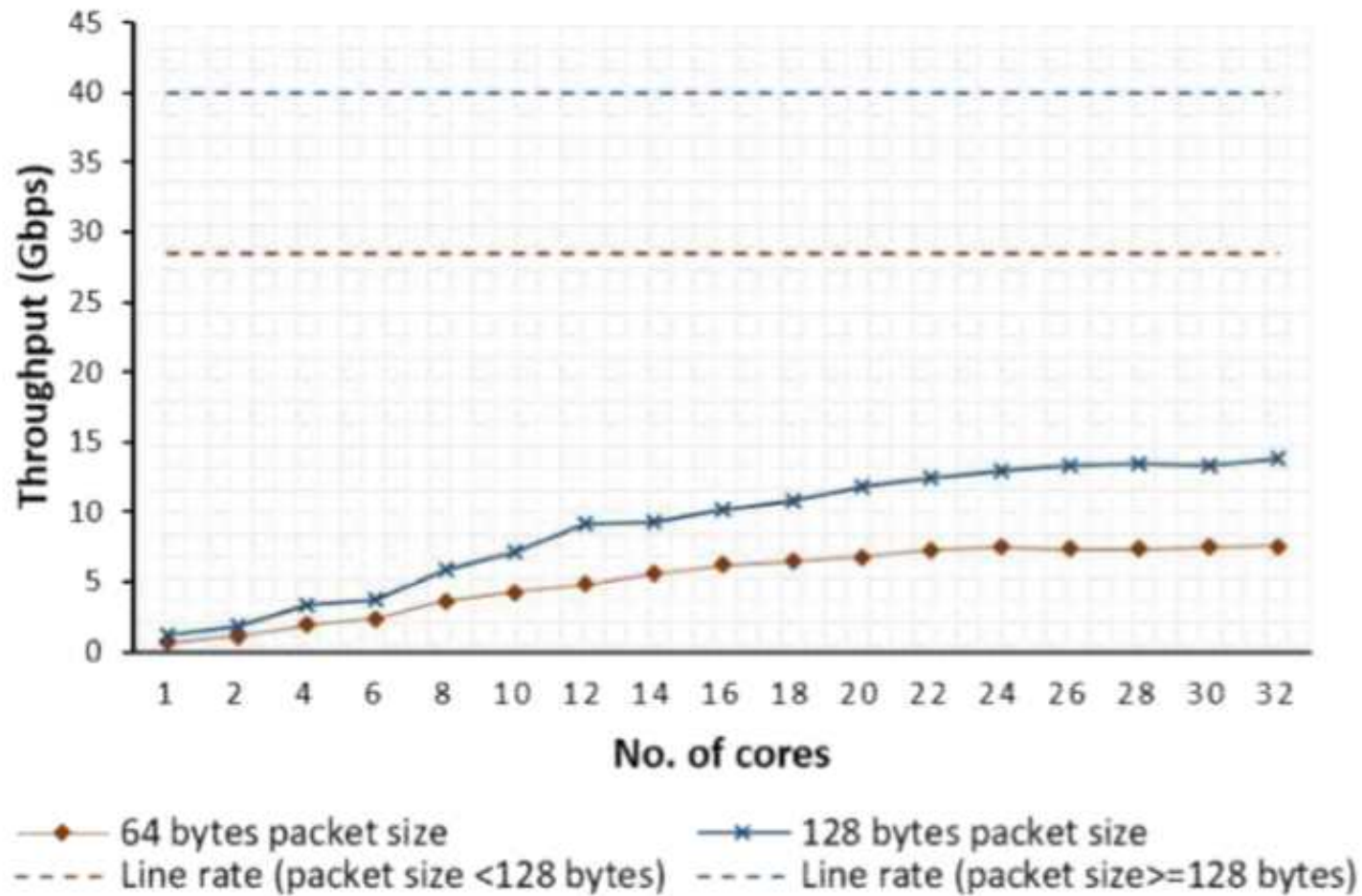


Figure 3: MOSBENCH results summary. Each bar shows the ratio of per-core throughput with 48 cores to throughput on one core, with 1.0 indicating perfect scalability. Each pair of bars corresponds to one application before and after our kernel and application modifications.

성능 그래프 (3)

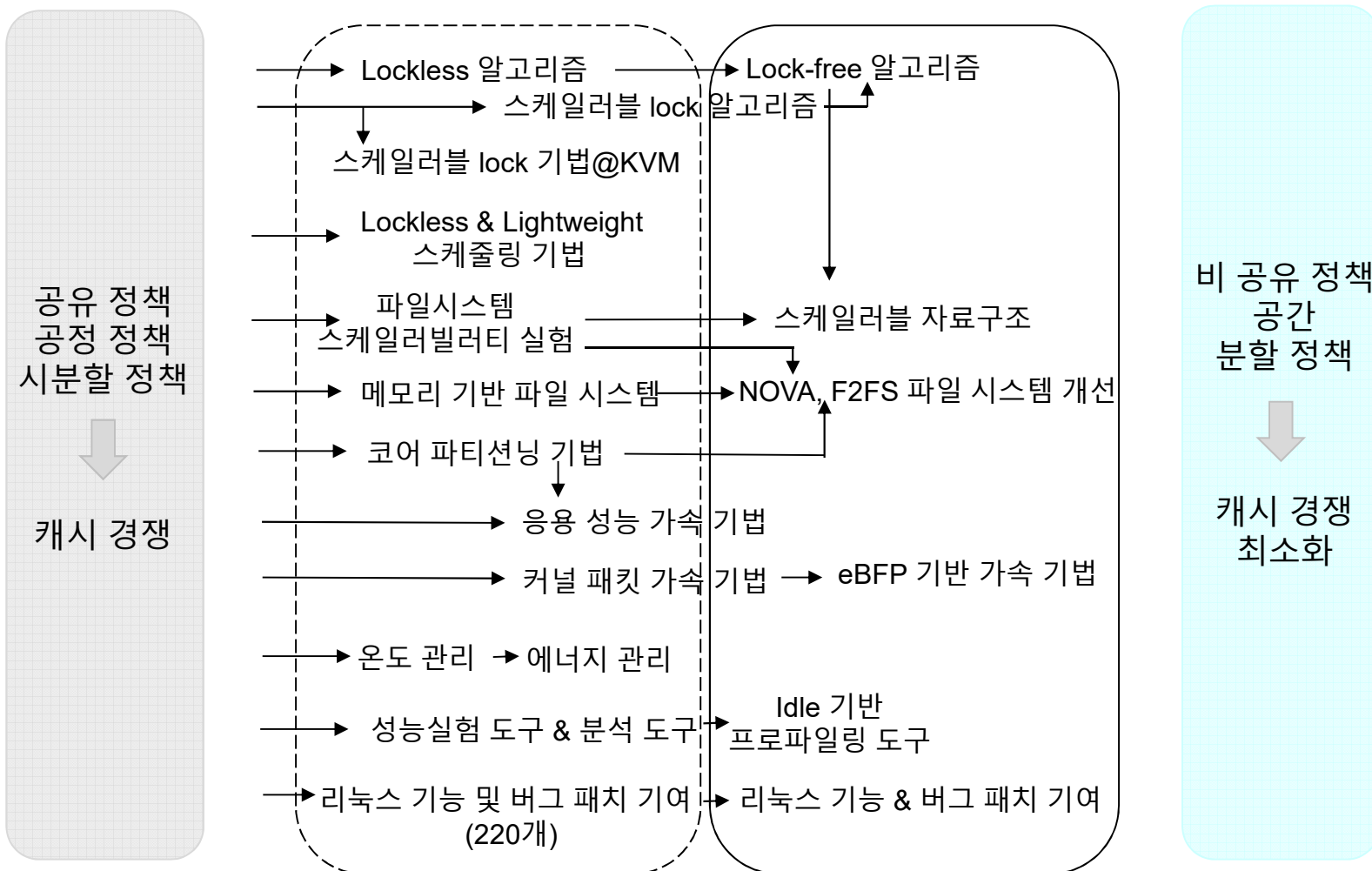
- 40G 네트워크



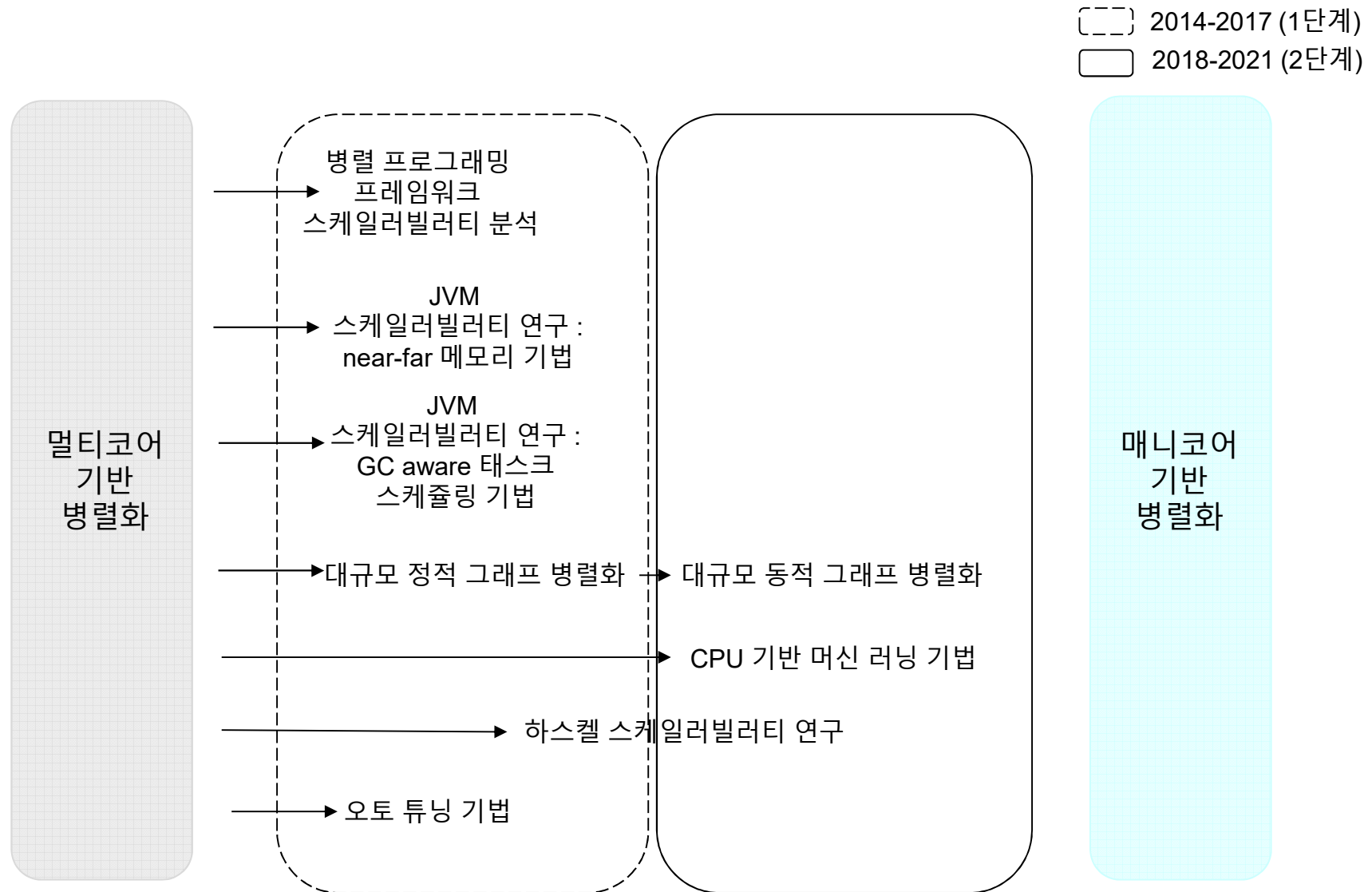
모노리틱 커널 스케일러빌리티 연구의 로드맵

[---] 2014-2017 (1단계)

[] 2018-2021 (2단계)

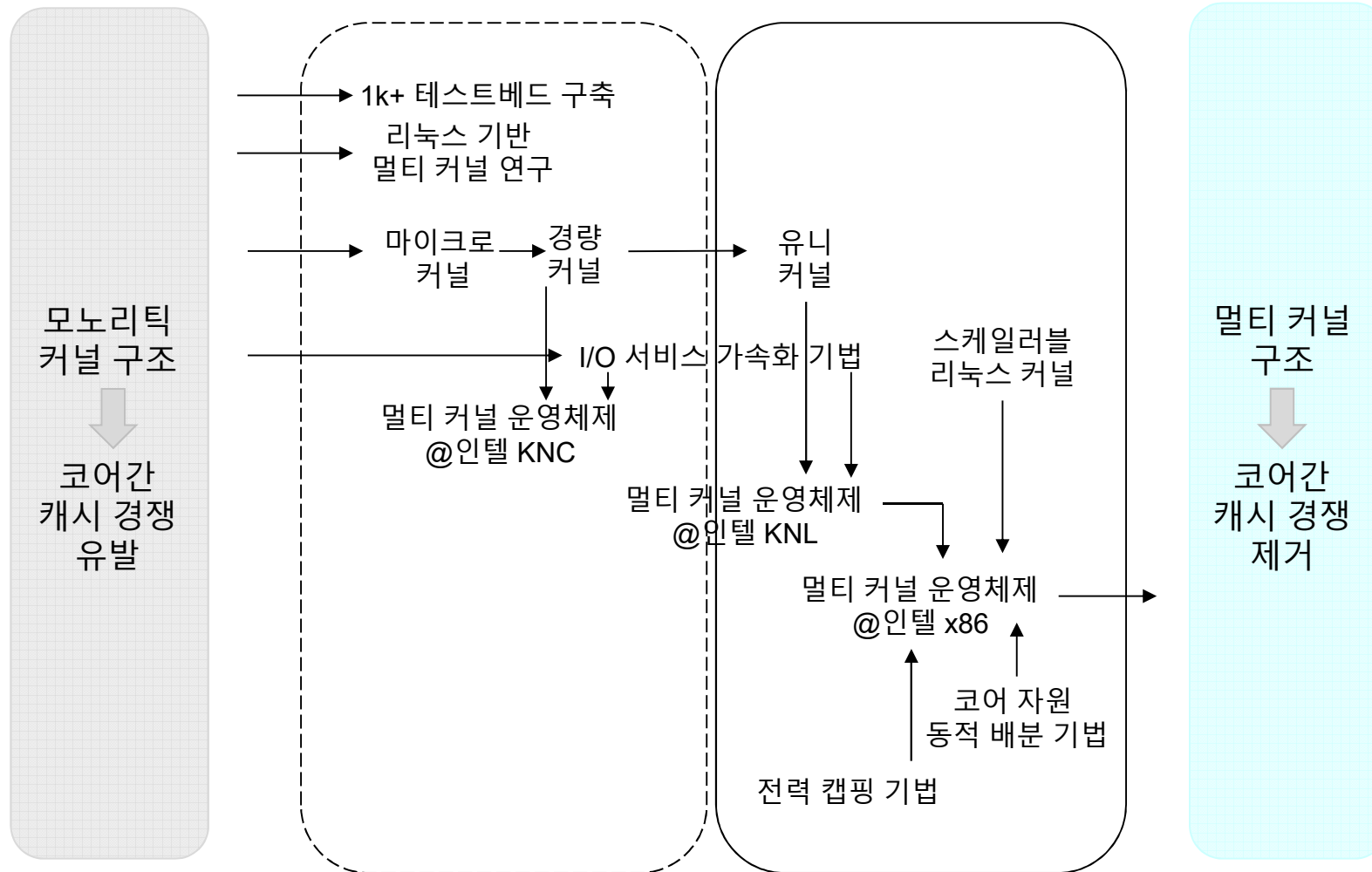


병렬화 연구의 로드맵



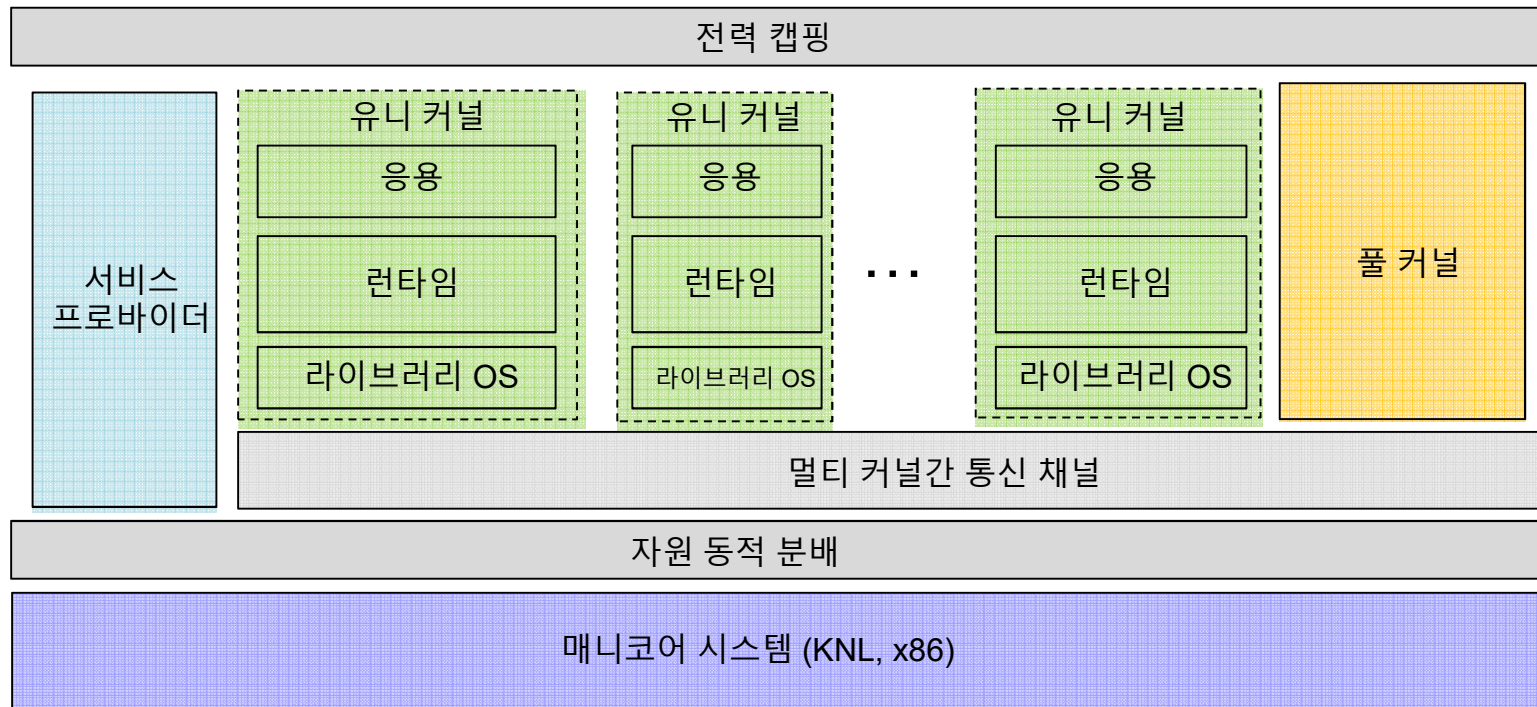
분할형 (멀티 커널) 운영체제 연구의 로드맵

[---] 2014-2017 (1단계)
[] 2018-2021 (2단계)



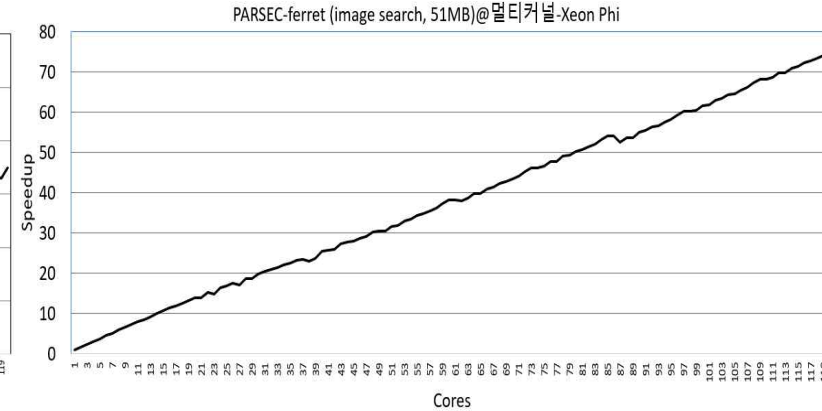
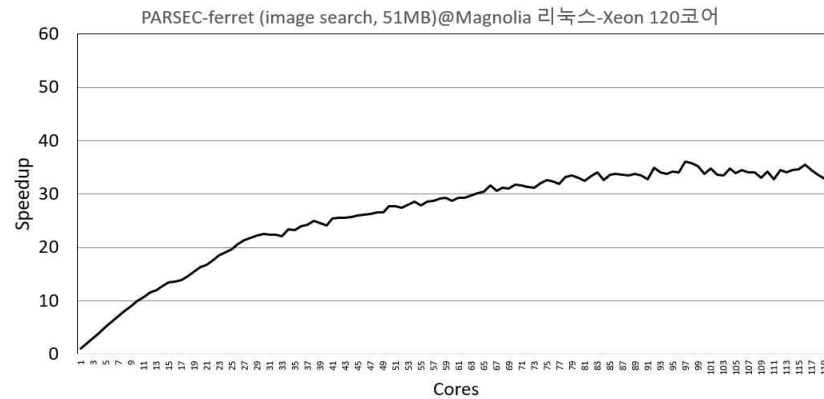
분할형(멀티커널) OS 연구

- 분할형 운영체제 구조도



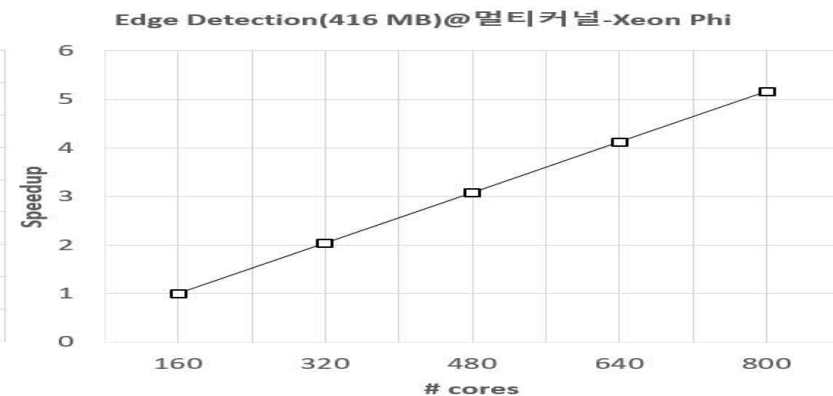
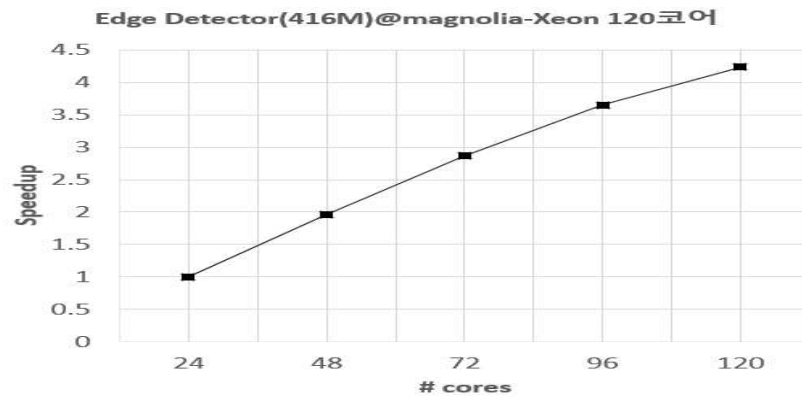
분할형(멀티커널) OS 연구

- 벤치마크/응용의 스케일러빌리티 비교 실험@4 Xeon Phi



<PARSEC ferret(이미지 유사성 검색, 51MB) 벤치마크 성능 비교 >

* PARSEC ferret : 데이터 공유가 많고, medium 병렬성의 워크로드



<이미지 검출 응용(416MB) 성능 비교 >

* 이미지 검출 응용 : 분할된 데이터 접근과 병렬성이 높은 워크로드

연구 결과물

Azalea-unikernel

Multikernel OS for manycore

● C 5 2 20

hybridF2FS

A variant of F2FS using

● C 0 0 0

sam

simple actor model f

● Haskell 0 0 0

qspinlockplus

oticketplus, qspinlockplus

● GPL-2.0 0 0 0

cst-locks

Implementation of C

● 0 0 0

mosaic

This repository is the main codebase for the Mosaic project

0 0 0 0 Updated on 31 Oct 2017

flsched

Feather-Like Scheduler

● C 0 2

sycalib

Dynamic core affinity f

0 0 0 0

iPocap

● C 0 0 0

Energysave

Linux kernel including featur

● C 0 0 0

unisan

UniSan: Proactive K

0 0 0 0

ivyproject_thermal

Thermal & Power management for Manycore OS

● C 0 0 0 0 Updated on 18 Feb 2016

latr

Latr: Lazy Translation Cohe

0 0 0 0

spm_managemen

● C++ 0 0 0

Azalea

Azealea is a new con system

● C 0 0 0

WASP

A workload-aware task sched framework

0 0 0 0

apisan

APISan: Sanitizing API Usages through Semantic Cross-Checking

0 0 0 0 Updated on 31 Oct 2017

deadline

0 0 0 0

Ramdisk_mq

● GPL-2.0 0 0 0

Ktune-1

Forked from ksy9164/Kt

● C 2 0 0

LDU

Forked from kmu-embedded/sca Linux Kernel Scalability Impr Performance - kernel source

● C 27,423 0 0

juxta

Juxta: Cross-checking Semantic Correctness for File Systems

0 0 0 0 Updated on 31 Oct 2017

eCS

0 0 0 0

parallelNOVA

A variant of NOVA files

● GPL-2.0 0 0 0

Magnolia

Magnolia is a Linux-t mechanism, energy efficiency and file systems.

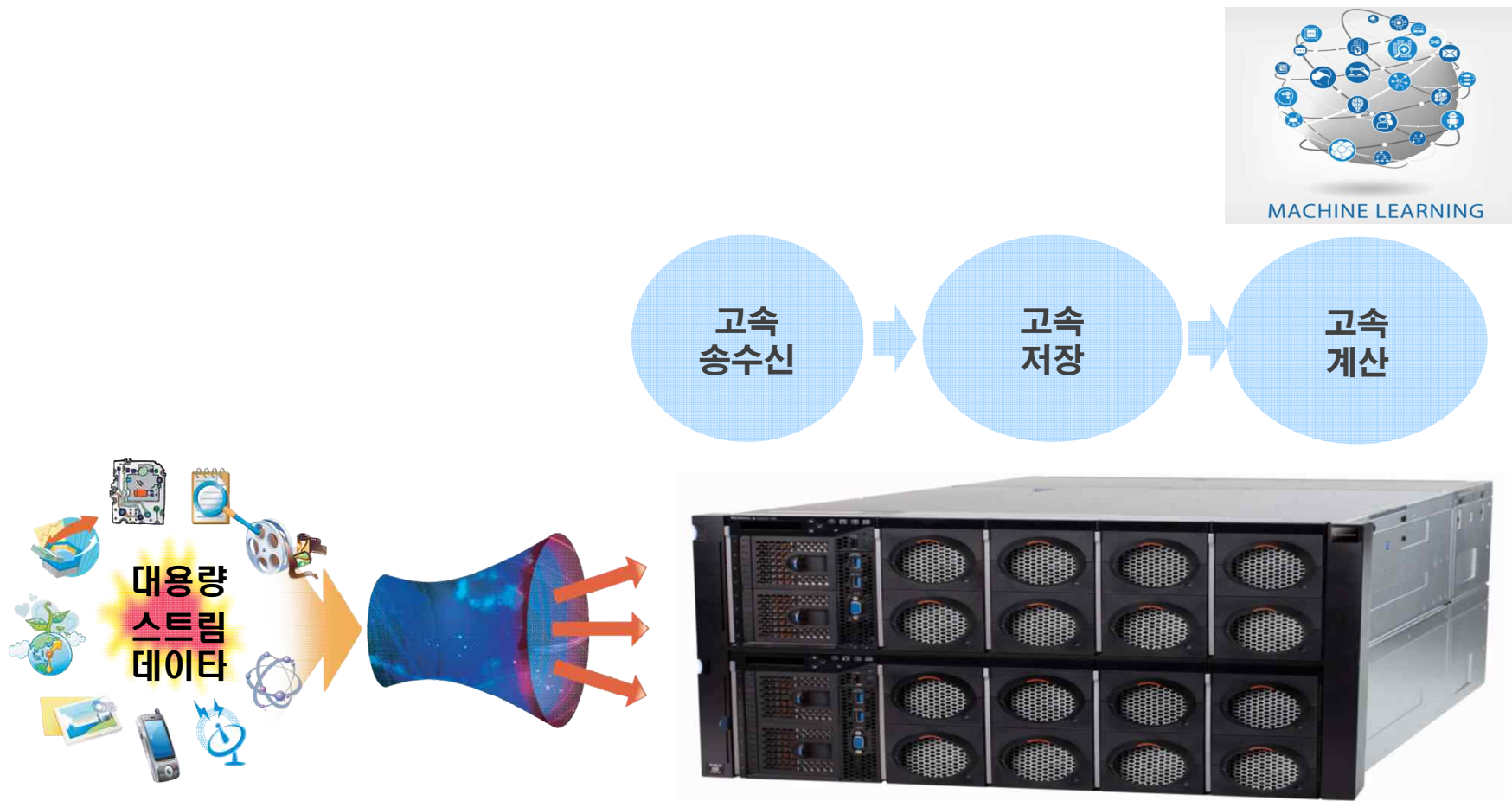
Ktune

fxmark

FxMark: Filesystem Multicore Scalability Benchmark

0 0 0 0 Updated on 31 Oct 2017


유스 케이스 (use case)

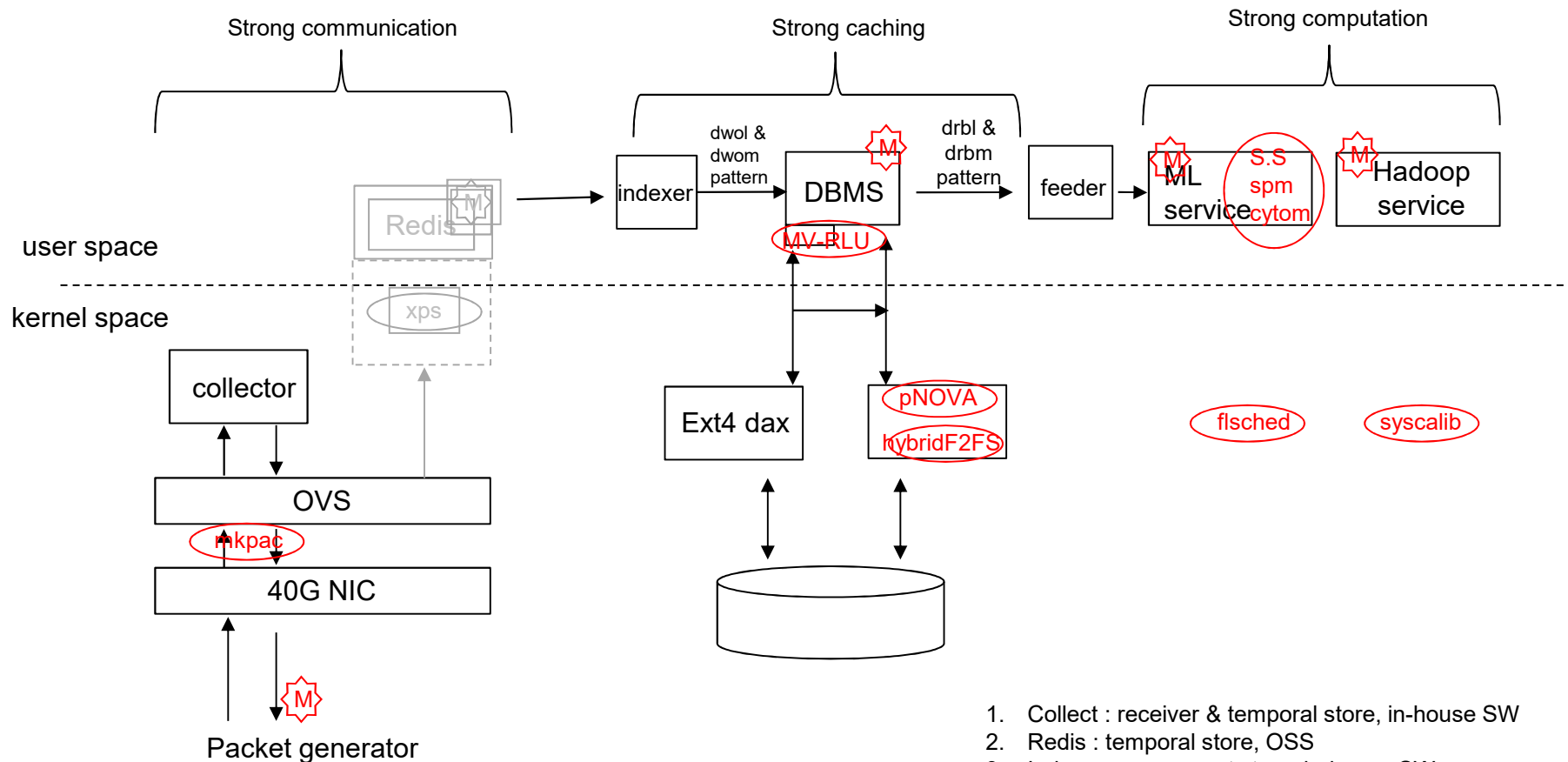


기념사합니다.



Detailed components of the 3C use case demo system

 Performance measurement points



1. Collect : receiver & temporal store, in-house SW
2. Redis : temporal store, OSS
3. Indexer : permanent store, in-house SW
4. DBMS : selecting among OSS DBMSs
5. Feeder : providing data to computation parts, in-house SW
6. dwol : overwrite a block in a private file
7. dwom : overwrite a private block in a shared file
8. drbl : read a block in a private file
9. drbm : read a private block in a shared file
10. Ext4 dax : ext4 direct access file system

- ❖ Youtube
- ❖ Open source relevant to 3C use case@<https://github.com/etri>
- ❖ Open issues : comparison target, # of cores, container?