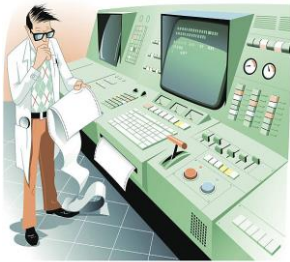# Operating System (OS) in the Era of AI

Xiao-Feng Li

xiaofeng.li@gmail.com

May 20, 2023

Evolution of Computing: User Interactions

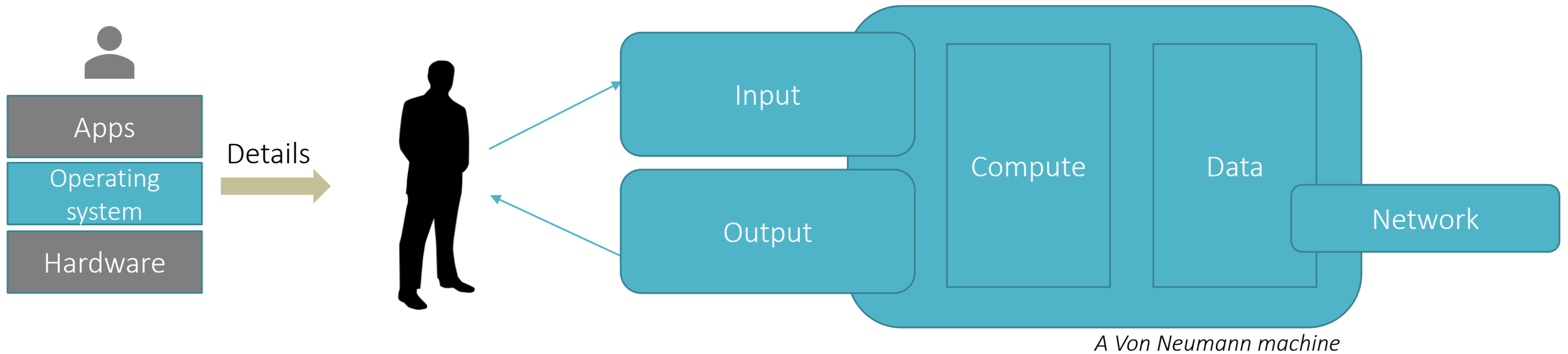Text and command   Mouse and GUI   Touch and mobile   Natural interaction

Transition

Users serve computer ➔ *Computers serve users*
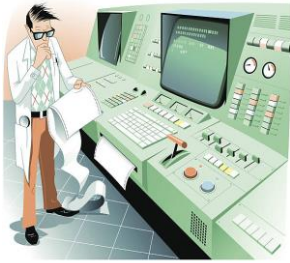
Core in Computing: Operating System

Role of Operating System

*Bridge user tasks and hardware (*such as *Windows, iOS, Android, Linux, etc.)*

Apps

Operating system

Hardware

Details

Input

Output

Compute

Data

Network

A Von Neumann machine

Evolution of Computing: Functionalities

Text and command

Mouse and GUI

Touch and mobile
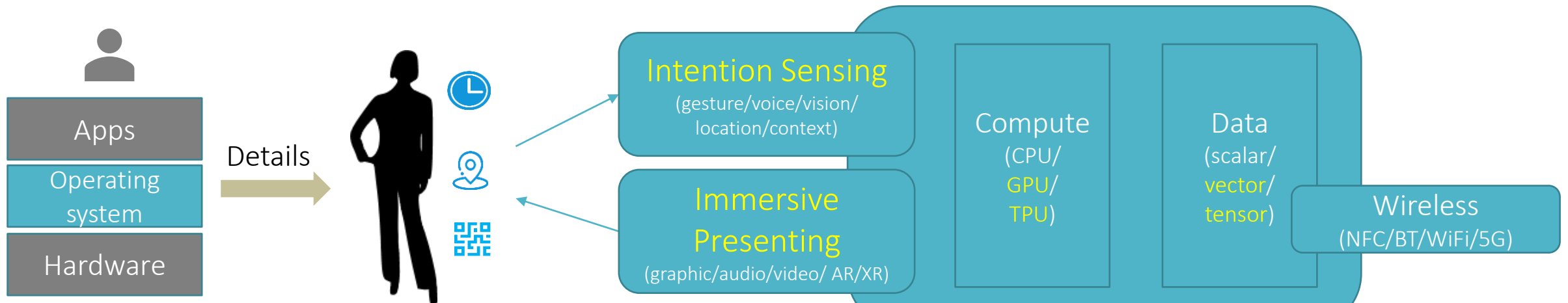
Natural interaction

Task and disk

Gfx and browsing
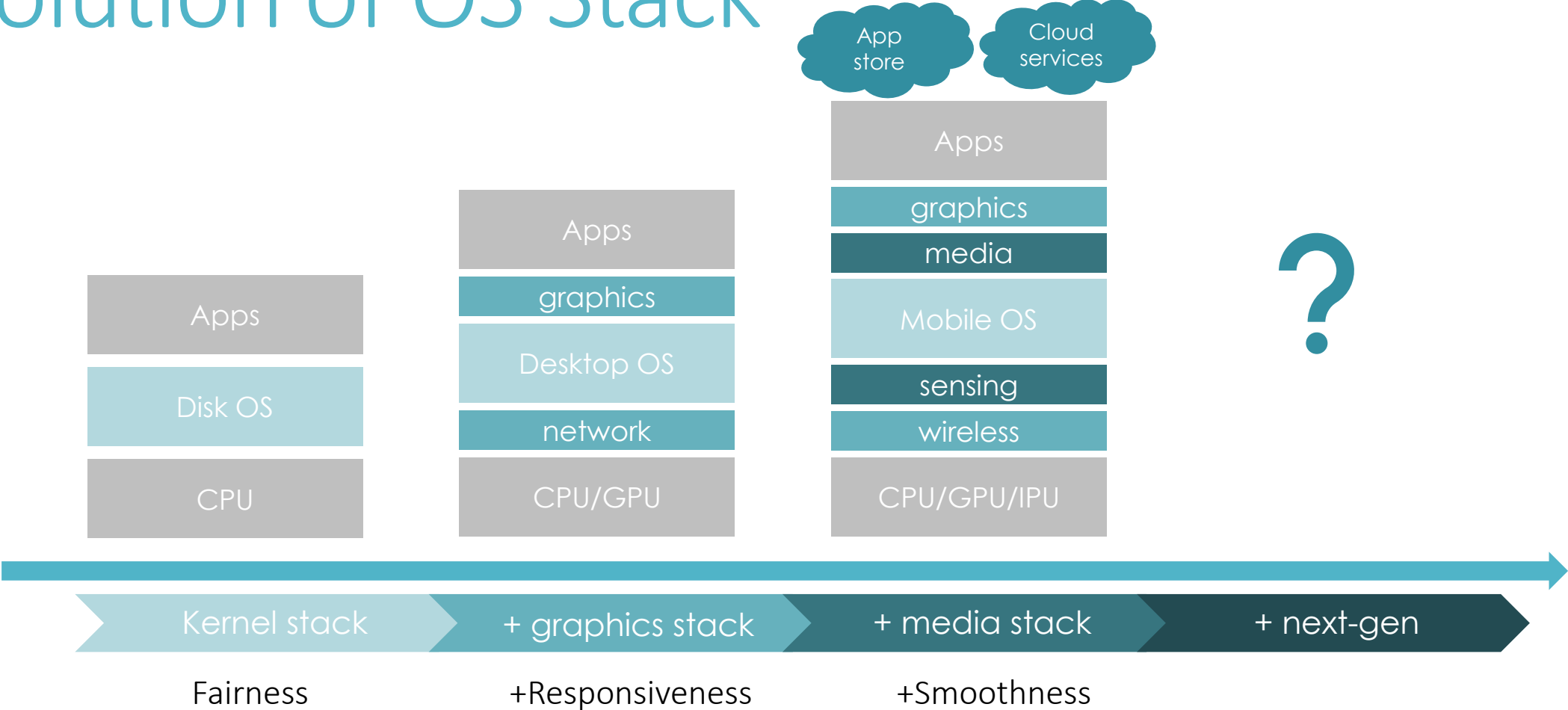
Media and comm.

Perception and prediction

Core in Computing: Current Status

Design goal of Operating System

hardware resource utilization ➔ *user value and experience*

Evolution of OS Stack

App store

Cloud services

Apps

graphics

media

Mobile OS

sensing

wireless

CPU/GPU/IPU

Apps

graphics

Desktop OS

network

CPU/GPU

Apps

Disk OS

CPU

?

Kernel stack → + graphics stack → + media stack → + next-gen

Fairness   +Responsiveness   +Smoothness

Thin client becomes thicker

Challenges to AI on End Devices

Transformer: 750x / 2 yrs
CV/NLP/Speech: 15x / 2 yrs
Moore's Law: 2x / 2 yrs
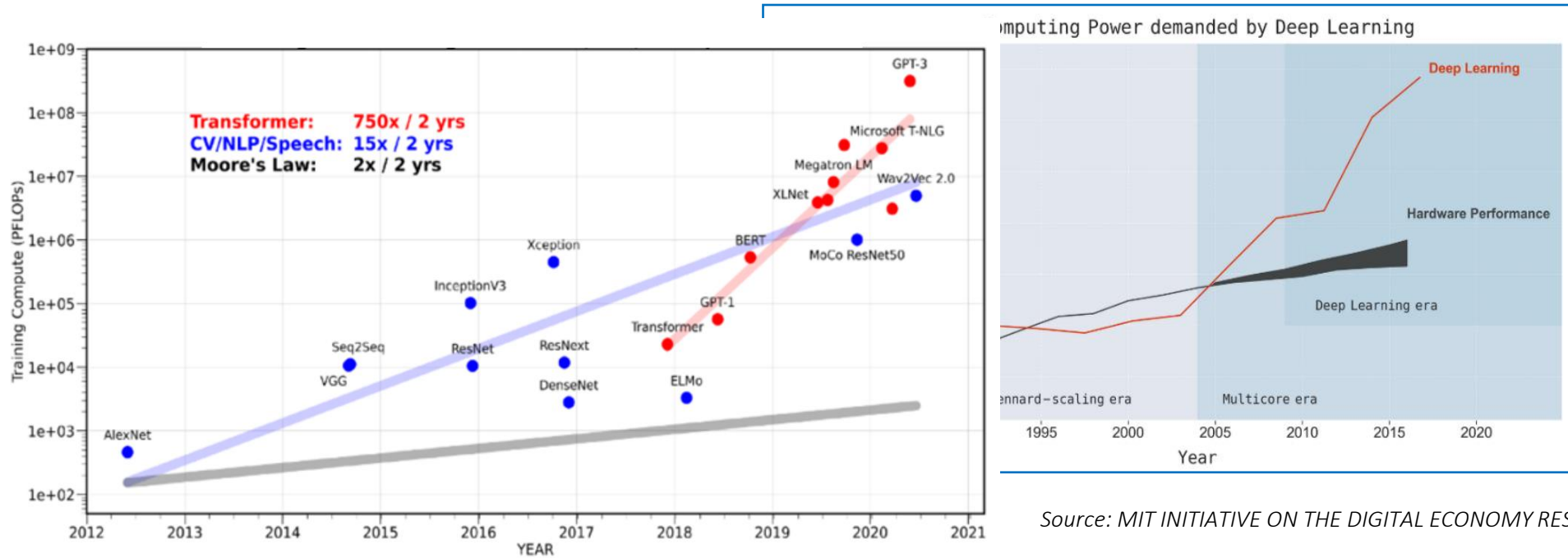
#Parameters:
GPT3.5: **175 billion**
GPT4.0: ~ **1 trillion**

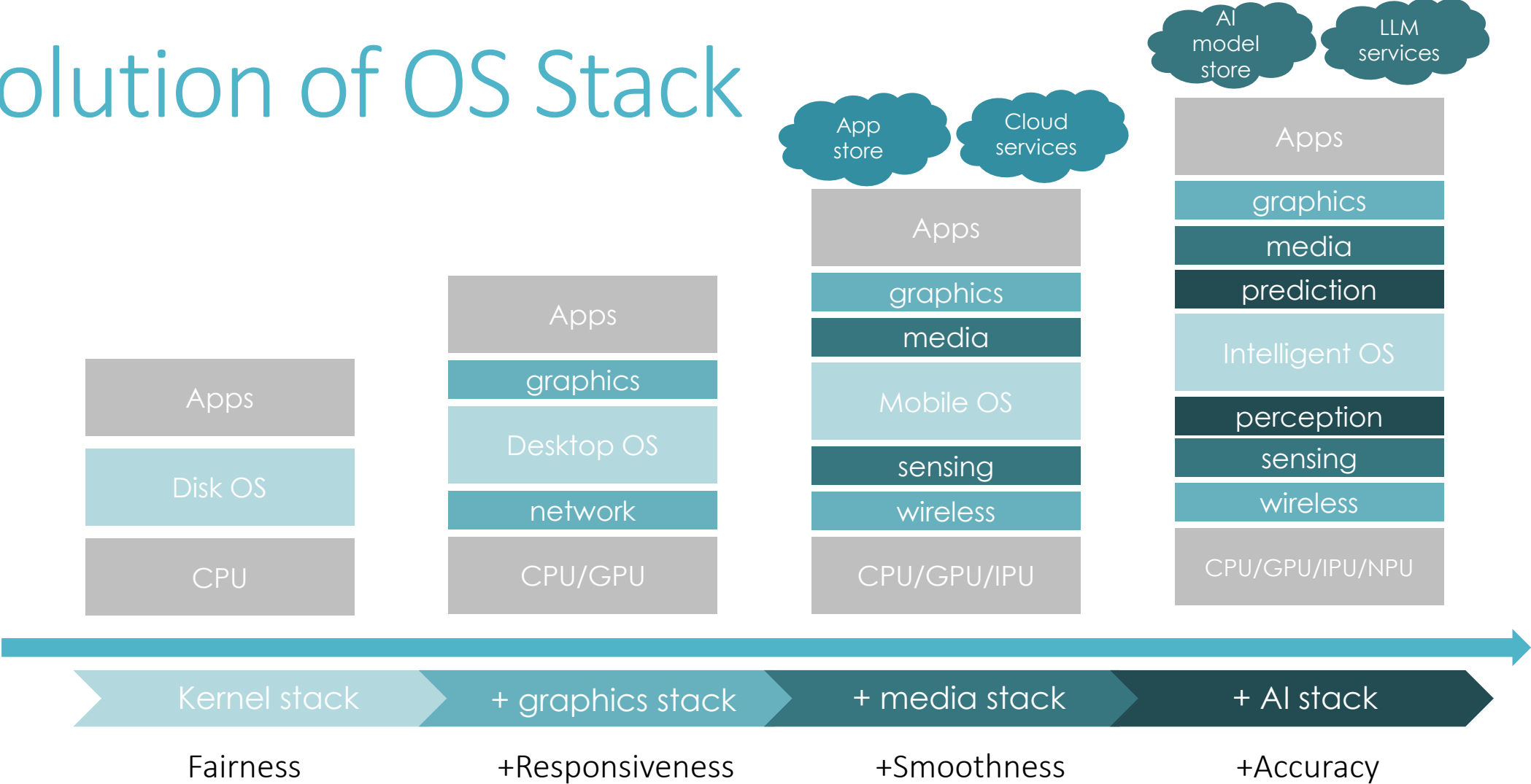*Source: MIT INITIATIVE ON THE DIGITAL ECONOMY RESEARCH BRIEF 2020 Vol. 4*

AI model size increases dramatically faster than chip capability

Evolution of OS Stack

Apps
Disk OS
CPU

Apps
graphics
Desktop OS
network
CPU/GPU

App store
Cloud services

Apps
graphics
media
Mobile OS
sensing
wireless
CPU/GPU/IPU

AI model store
LLM services

Apps
graphics
media
prediction
Intelligent OS
perception
sensing
wireless
CPU/GPU/IPU/NPU

Kernel stack → + graphics stack → + media stack → + AI stack

Fairness   +Responsiveness   +Smoothness   +Accuracy

Thin client becomes *thicker* *and* *thinker*

Opportunities of Next-Gen OS

Design of operating system
App-centered ➡ *AI-centered*
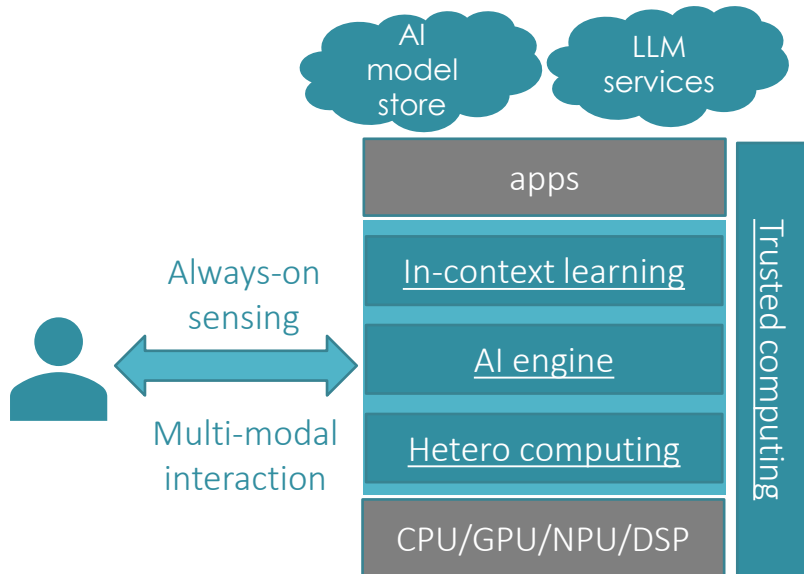
- Understanding user needs

  Low power always-on sensing and *end-to-end trusted computing*

- Execution of AI tasks

  AI model inference acceleration and *heterogenous computing*

- Intuitive user interaction

  Multi-modal interaction and *smart in-context learning*

Thanks!