LPW(LP WIDE IO) Introduction

Feb. 2024 | Samsung Memory

Confidential

Caution for Document License

Agendas

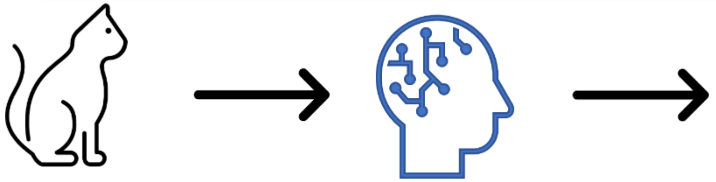
1) Market Introduction
   - ◻ AI Market Trend
   - ◻ On-Device Generative AI Analysis and Memory Roadmap

2) AI Memory Solution for Mobile
   - ◻ LPW(LPDDR Wide IO) Concept Introduction
   - ◻ LPW(LPDDR Wide IO) Architecture
   - ◻ LPW(LPDDR Wide IO) Packaging

Market Trends for Artificial Intelligence

1. **[~2023] New Models, Learning new capability from existing data → Training**
   1) Objective: Build Deep Learning Framework (e.g. GPT, PaLM, LLaMA)
2. **[2023~] Introducing real-life applications, Applying capability to new data → Inference**
   1) Objective: Real life Apps or Services (e.g. ChatGPT, Bard)
      - Generative Models & On-device support

Training

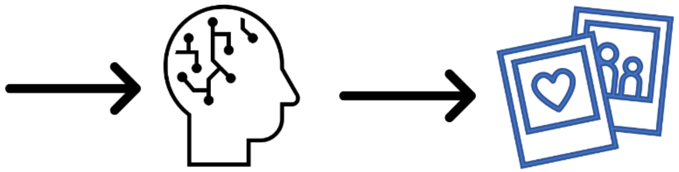


Cat (O)
Dog (X)

Model Training
Server / Datacenter
Maximum Performance

Inference

"Give me a Picture of Cat"



Optimized Model
On-device
Power Efficient

Mobile Market Driven by On-device Generative AI

1. **Mobile: Simpler queries & less accuracy → ASR for Voice assistant, light-weight photo editing for social media**
    1) Large Language Model for Mobile : Compromised model for target UX
2. **LP Mem. Solution : LP5X (D1b 9.6Gbps → 10.8Gbps) → Next Gen and Pathfinding (LP6 & LP Wide IO, LP-PIM)**

(M.P. based / Left edge & center(circle) aligned)

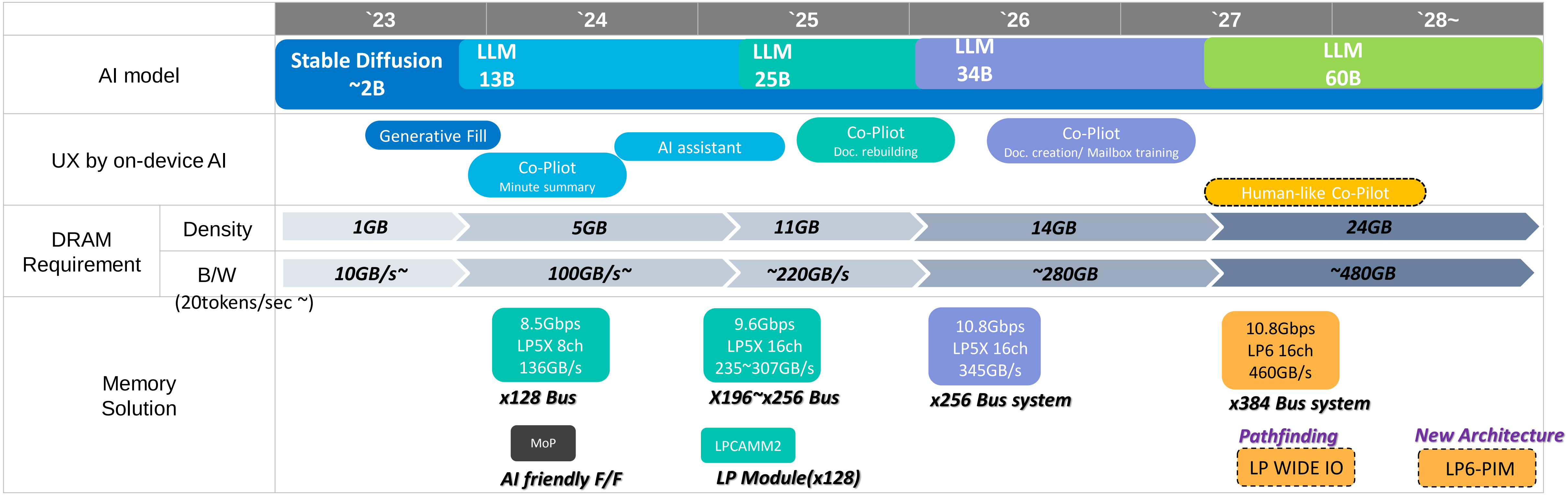
		`23	`24	`25	`26	`27	`28~
AI model		Stable Diffusion ~2B	Light LLM ~7B	LLM 13B	LLM 20B	LLM 34B	
UX by on-device AI		Image create	ASR summary, Short NLP	AI Assistant	Human-like AI Assistant		
DRAM Requirement*	Density	1GB	2.7GB	5GB	7.6GB	13GB	
	B/W	10GB/s~	27~54GB/s	50~100GB/s	76~152GB/s	130~260GBps	
Memory Solution			8.5Gbps LP5X 68GB/s	9.6Gbps LP5X 76GB/s	10.8Gbps LP5X 86GB/s	* Next Gen Memory: Next Gen JEDEC: 10.8Gbps LP6 115GB/s; New Architecture: LP6 PIM; LP WIDE IO 204GB/s Pathfinding	

* DRAM requirements assumptions : ~60% capacity compressions / 10~20tokens/sec AI performance

Confidential

PC Market Driven by On-device Generative AI

1. **AI-enabled PC is expected to be deployed for commercial area first and then gradually spread to consumer area**
    1) Commercial : Productivity & Creativity (e.g. Paperwork automation, Auto code creation, AI generated meeting minutes)
    2) Consumer : Convenience & Entertainment (e.g. Personal assistant, Real-time avatars, AI enhanced gaming)
2. **Multi-Channel system(x128↑) for sufficient BW and High-Density solution for utilizing larger model**



- LLM Memory requirements assumptions
  - 60% Memory footprint reduction : Quantization(Int4) + HW Compression/Accelerating logic

Post LP5X Memory Solutions for Generative AI

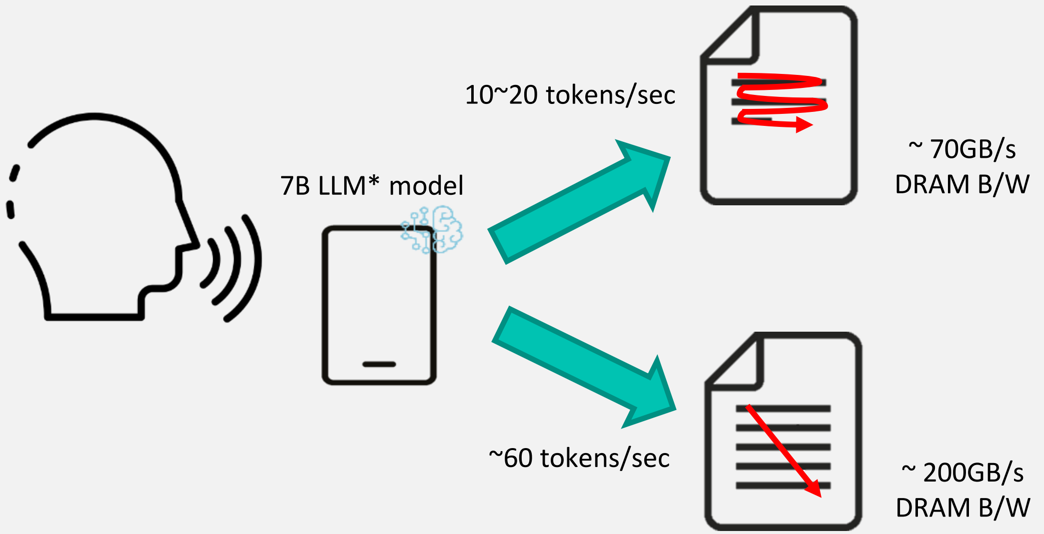
1. On-Device Gen AI (Memory bounded UX) → Higher BW & Lower power LP Memory sub-system is crucial

1) LPDDR6 : ~120GB/s BW, Lower power feature(VDD2D 0.875v / Efficiency mode) JEDEC standardization
→ Flagship SOCs will support LPDDR6 targeting '26.2H mass production

2) LP Wide IO: 204.8GB/s Total B/W, Leveraging Wide IO architecture to attain lower power (lower pin speed, 3.2Gbps)
→ Need to explore packaging solutions (e.g Wide IO level vertical wire-bonding & Fine Pitch bonding option)

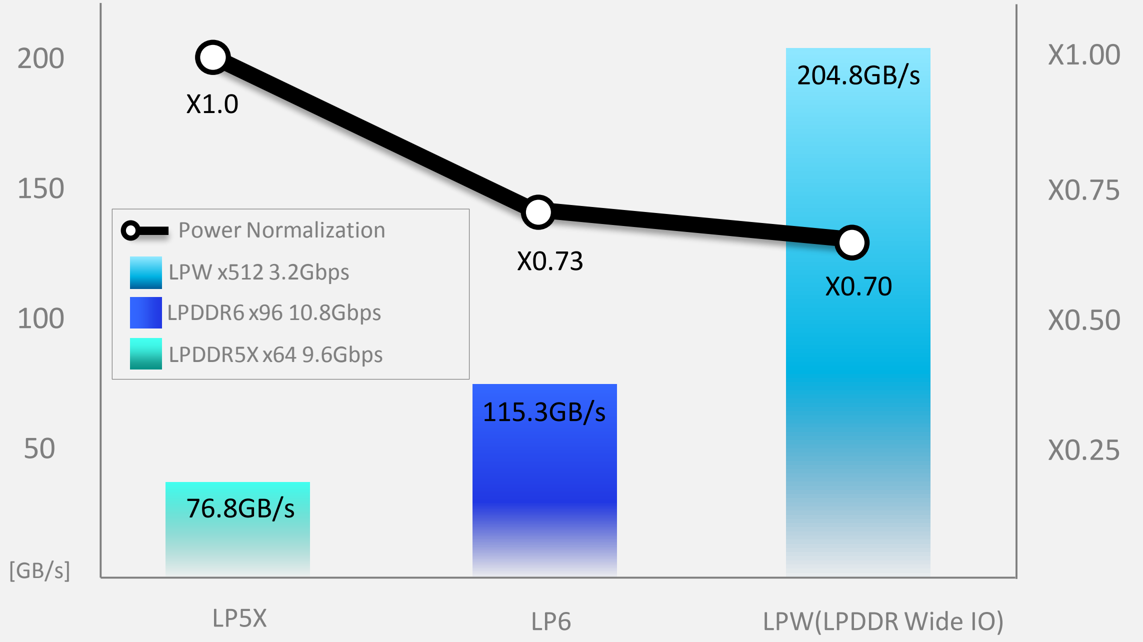
Higher Memory B/W Requirements

✓ Paragraph-by-paragraph fast reading and response
· Current GPT engine token generation speed : 4~6 words/sec



(* Large Language Model)

Power Efficiency& Total Bandwidth per PKG

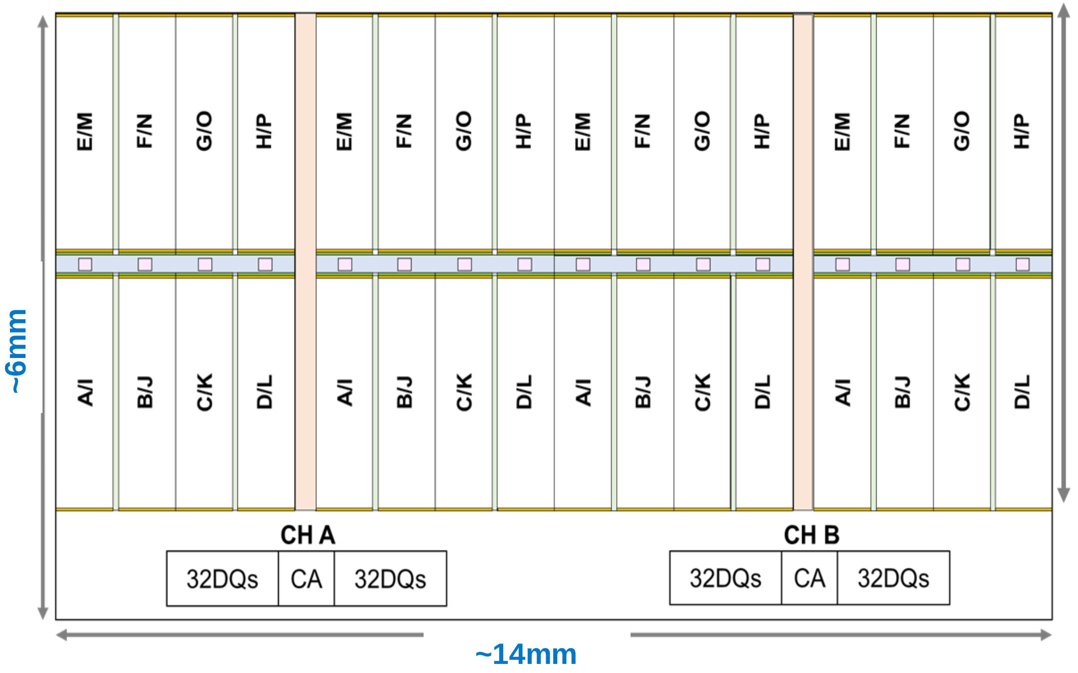


(* Compared with same process node)

LPW(LP Wide IO) introduction

1. **Samsung is now reviewing a large-capacity LP wide-IO as main memory solution**
   1) Capacity : 24Gb, Peak BW : 51.2GB/sec, Power efficiency : ~1.9pJ/b and DQ Count : 128 w/Max. 3.2Gbps

Die Arch



✓**2Channels**
- 2 p-ch per channel

✓**16 banks per pseudo channel**

✓**Open-page w/explicit activates**

✓**Data IO**
- Total 128DQs (32DQs per pseudo channel)

Confidential

LPW(LP Wide-IO) : Die architecture

1. **LPW die architecture is similar to LLW**
   But larger capacity considering main memory solution with stacking
   1) Other key differences : Channel count IO, Data rate, Core timing and use of explicit activates

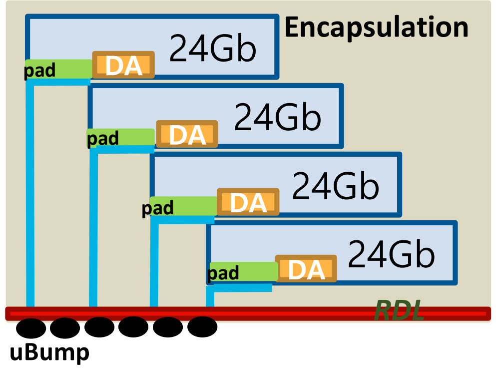
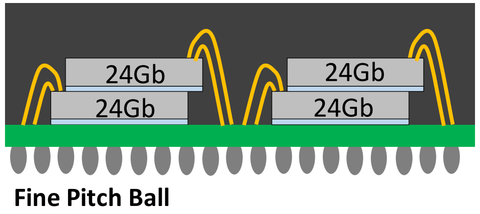
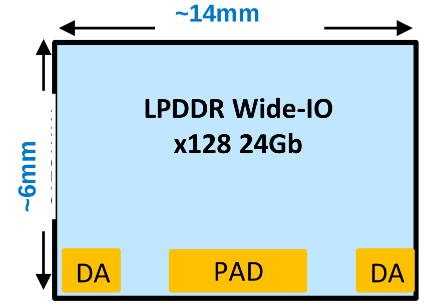
	LLW (Low Latency Wide IO)	LPW(LP Wide IO)	LPDDR6
Capacity and BW	1Gb, 128GB/s	24Gb, 51.2GB/s	16Gb, 28.8GB/s
Organization	4 channel 2 pseudo channels (pCH) per channel 8 banks per pCH (No bank groups) x64 DQ per pCH w/Burst length 8, 16, 32 → Min 64B Closed-page operation w/Implicit activates	2 channel 2pseudo channels (pCH) per channel 16 banks per pCH (no Bank groups) x32 DQ per pCH w/burst length 16 and 32 → Min 64B (* BL8 can be considered for Min 32B) Open-page operation w/explicit activates	2 channel 2 Sub channels (Sub-CH) per die 16 banks per Sub-CH (4Banks/4BG) x12 DQ per Sub-CH w/burst length 24 and 48 → Min 32B Open-page operation w/explicit activates
Signaling and clocking	DDR Signaling for DQ and CA 96 Mbps, 1, 2Gbps data rates 1 diff. clock per channel 1 diff. DQS per x32 DQ	DDR Signaling for DQ and CA 0.8, 1.6, 3.2Gbps data rates 1 diff. clock per channel 1 diff. DQS per x32 DQ	DDR Signaling for DQ and CA 10.8Gbps data rates (1st target) 1 diff. clock per Sub-channel 1 diff. DQS per Sub-channel
Critical Core Timing	tRCmin=28/32ns, Latency=30ns	tRCmin=60ns, Latency(tRCDr + RL + BL16) < 42 ns	tRCmin=60ns
Refresh	Only all-bank refresh per pseudo channel	Both per-bank and all-bank refresh	Per dual-bank and all-bank refresh
Calibrations	Only at cold boot + background periodic ZQC	Only at cold boot + background periodic ZQC	background periodic ZQC
Testability	Direct Access Port, Boundary Scan	Direct Access Port, Boundary Scan	N/A
Reparability	Lane Repair, Post-Package Repair	Lane Repair, Post-Package Repair	Post-Package Repair,

Case study for LPW PKG options : <u>VWB</u> +RDL, Fine-pitch PKG, Wafer Biz

**Vertical Wire Bonding*

1. Currently, LPW is pathfinding status & now reviewing technical feasibility on packaging options

1) Case : 1. VWB + Wafer level RDL w/new equipment, 2. Fine-Pitch PKG w/ existing infra, 3. Wafer Biz

<table>
<tr><th colspan="3">Case study</th><th>1. VWB + RDL</th><th>2. Fine-pitch PKG</th><th>3. Wafer Biz</th></tr>
<tr><td>LP Wide-IO</td><td colspan="2">Concept</td><td>
</td><td>
</td><td>
</td></tr>
<tr><td rowspan="4">Tech. feasibility</td><td colspan="2">Chip die</td><td>LPW architecture</td><td>←</td><td></td></tr>
<tr><td rowspan="3">PKG</td><td>Stacking</td><td>New solution
-VWB@4H stacking</td><td>D-DDP structure
- x256/512PKG @wire bonding</td><td>N/A</td></tr>
<tr><td>Pad Pitch</td><td>60+@µm</td><td>80µm
-Staggered</td><td>.</td></tr>
<tr><td>Size</td><td>Reviewing</td><td>~16.0mm x ~14.0mm
@0.27ball pitch assumption</td><td>.</td></tr>
<tr><td colspan="3">Risk point</td><td>VWB feasibility
Testability @DA Pad is consideration</td><td>PKG Size increase
• Reviewing 4H stack option to minimize Y-axis (Tentative size : ~16mm x ~9mm)</td><td>.</td></tr>
</table>

A journey shared takes us beyond