MAKING KGD SILICON WORK IN YOUR SUPPLY CHAIN

DAVID GREENLAW
VP Product Engineering

16-SEP-2020
TOPICS

• CMOS scaling over time
• Growth in compute demand remains exponential
• A100 and DGX-A100
• Types of manufacturing relationships
• Rise of silicon foundries is instructive
• What’s missing in the Adv Pkg discussion
• Summary
# 30 YEARS - SIMPLIFIED VIEW

<table>
<thead>
<tr>
<th>1990s</th>
<th>2000s</th>
<th>2010s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moore’s Law Dominant</td>
<td>Moore’s Law Difficult</td>
<td>Moore’s Law Slowing</td>
</tr>
<tr>
<td>‘Heyday’ of CMOS Scaling</td>
<td>Increased Scaling R&amp;D $</td>
<td>Less Scaling, more stacking</td>
</tr>
<tr>
<td>Timeframe</td>
<td>Limiter</td>
<td>Problem</td>
</tr>
<tr>
<td>-----------</td>
<td>-------------------------------------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>1980's</td>
<td>Multi-Metal Chip Planarity</td>
<td>Interconnect Scaling</td>
</tr>
<tr>
<td>1990's</td>
<td>Multi-Metal Interconnect Capacitance</td>
<td></td>
</tr>
<tr>
<td>1990's</td>
<td>Metal Etch</td>
<td></td>
</tr>
<tr>
<td>2000's</td>
<td>Gate Oxide leakage</td>
<td>Transistor Scaling</td>
</tr>
<tr>
<td>2000's</td>
<td>Poly Gate Depletion</td>
<td></td>
</tr>
<tr>
<td>2000's</td>
<td>Short Channel Control</td>
<td></td>
</tr>
<tr>
<td>2010's</td>
<td>sub-193nm transmissivity through lenses</td>
<td>Cost of Patterning</td>
</tr>
</tbody>
</table>
SCALING NOW FOR WHAT YOU NEED MOST

Moore’s Law used to improve all 3 together…

… so which was “most important” didn’t matter

Many Cost Sensitive Products

Cost

Density

CPUs …

Serial Perf / Watt

GPUs and other ‘inherently parallel’ processors

“If you could only pick ONE of these for a ~10% improvement …”
ACCELERATED COMPUTING
1000X EVERY 10 YEARS

- GPU PERFORMANCE: 1.1x per year
- Single-threaded CPU: 1.5x per year
EXPONENTIAL GROWTH IN COMPUTING DEMAND

DATA SIZE GROWING

Zettabytes Generated

AI RESEARCH GROWING

Papers Submitted to NeurIPS & CVPR

AI MODEL COMPLEXITY GROWING

Petaflop-Days (Training)

Source: IDC, GitHub, and OpenAI / NVIDIA
NVIDIA A100
TSMC 7nm | HBM2 – 1.6 Terabytes per Second | 3D Chip Stack

54 BILLION XTORS
3rd GEN TENSOR CORES
SPARSITY ACCELERATION
MIG
3rd GEN NVLINK & NVSWITCH
NVIDIA DGX A100
3rd Generation
Integrated AI System

5 PetaFLOPS of Performance in a Single Node

Unified System for End-to-End Data Science and AI

Fully Accelerated Stacks — Spark 3.0, RAPIDS, TensorFlow, PyTorch, Triton

Elastic Scale-Up or Scale-Out Computing

High Scalability with Mellanox Networking
NVIDIA DGX A100
3rd Generation
Integrated AI System

5 PetaFLOPS of Performance in a Single Node

- 9x Mellanox ConnectX-6 VPI
- 200 Gb/s Network Interface
- Dual 64-core AMD Rome CPU
- 1 TB RAM
- 8x NVIDIA A100 GPUs
- 6x NVIDIA NVSwitches
- 15 TB Gen4 NVMe SSD
NVIDIA DGX A100
3RD GENERATION
INTEGRATED AI SYSTEM

5 PetaFLOPS of Performance in a Single Node

150X  AI Compute
40X   Memory Bandwidth
40X   IO Bandwidth

Compared to High-End CPU server

Available Now at $199K
KGD & ADVANCED PACKAGING: TWO PROBLEMS

What everyone talks about:

- Stacking and Advanced Packaging Technologies
- Compelling Business Cases

Mostly Missing:
- Adopting wafer fab Internal Working Culture onto the assembly floors
  - SPC
  - FMEA
  - ‘continuous improvement’ all the time ...

Mostly Missing:
- Adopting foundry-like External Working Culture
  - Customer Service ...
MANUFACTURING RELATIONSHIPS

IDM MODEL
IDM: Owns all Mfg data
Design Data
Silicon Data
Tester Data
Simplest for mfg because all data within one company

FABLESS/FOUNDRY MODEL
Fabless Company:
Design Data
Tester Data
Interactions to connect statistical tester data to statistical silicon data
Silicon Foundry:
Silicon Data

CUSTOMER/SUPPLIER MODEL
Customer Company:
Small pile of failing die
No data ...
Interactions inefficient because we have no statistical data about anything
Memory Supplier:
Design Data
Silicon Data
Tester Data

Then: original doubters of the Foundry model could not imagine solving complex problems between 2 companies
Now: Foundry model works. KGD silicon requires solving complex problems between multiple companies
25 YEARS AGO

Silicon Foundry

- Run the Fab
- Technology Development
- Customer Focus

DRAM Supplier

- Run the Fab
- Technology Development
- Customer Focus

strong from the beginning
weak in the beginning
Slowing Moore’s Law Scaling means: “Everyone has to be good at everything”

Corollary: companies moving from Packaging to Advanced Packaging will work on all three...
CUSTOMER FOCUS - WHAT, EXACTLY?

**Means this**

- See your customer’s success as key to your own success
- “Own the whole problem”, drive to closure
- Internal leverage from Customer Eng to internal teams...
- Engineers want to work on solving customer problems ... connected to their career success

**Does NOT mean**

- have to always ‘be nice’ to the customer ...
- measured in # ppt slides with customer logo
- measured in # teleconf hours/month ...
- etc
SUMMARY

- Silicon scaling remains limited by cost of patterning
- But exponential growth of compute demand continues
- KGD silicon and Advanced Packaging a part of the solution

Adv Pkg work: ‘Technology’ and ‘Business’ necessary, but insufficient
  - previous rise of the Foundry/Fabless model is instructive
  - Working Culture is also a key to success:
  - successful Advanced Packaging factories behave more like wafer fabs: SPC, ‘continuous improvement’, etc
  - successful KGD and Packaging suppliers behave more like Foundries: Customer Service, working across supply chain
Thank you sponsors!
ASE - Expanding Our Role in the Electronics Ecosystem

Application-Specific Solutions

OSAT

EMS

Semiconductor

Silicon / Module Integrator

System OEM

Content/Service

User

The Industry’s most comprehensive toolbox
Providing a complete value chain solution
A global leader in the ATE industry with a WW installed base of over 30,000 systems.

Our nanotechnology products support leading-edge semiconductor processes at the 1Xnm node.

Our diverse workforce includes 5,500 employees from 50 countries.

Eco-friendly policies emphasize reduction of our carbon footprint.

2018 Global Technology Leader by Thomson Reuters.

Innovating in the measurement arena for 60+ years.

A VLSIresearch 10 BEST supplier for 32 consecutive years.
# Carrier Solutions for Known Good

<table>
<thead>
<tr>
<th>Vacuum Release Carriers</th>
<th>Gel-Box • Gel-Tray • Gel-Slide Carriers</th>
<th>NEW Carrier Films</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pocketless Trays for Automated KGD Handling</td>
<td>Carriers for Manual KGD Handling</td>
<td>• Reconstituted Known Good Wafer Handling</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Universal Carrier</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Custom Constructions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Low Tack, Low Residue</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Textured Available</td>
</tr>
</tbody>
</table>

www.gelpak.com
1-888-621-4147

A division of DELPHON
COPYRIGHT NOTICE

This presentation in this publication was presented at the Known Good Die (KGD) Workshop 2020. The content reflects the opinion of the author(s) and their respective companies. The inclusion of presentations in this publication does not constitute an endorsement by MEPTEC, IMAPS, or the sponsors.

There is no copyright protection claimed by this publication. However, each presentation is the work of the authors and their respective companies and may contain copyrighted material. As such, it is strongly encouraged that any use reflect proper acknowledgement to the appropriate source. Any questions regarding the use of any materials presented should be directed to the author(s) or their companies.

www.kgdworkshop.org
www.meptec.org