Intel 1GB Chiplet Feasibility

Follow-up items after 11/30/22 call

This summary contains AP Memory proprietary and confidential information, including patented and patent-pending technologies. Disclosed to Intel Corp. under M-RUNDA.

AR from 11/30/22 call

- Die size vs bandwidth
- Die size vs latency
- Wafer bow limit for HB

Die size vs Bandwidth

- Corrected proper accounting of bandwidth
  - The wording in 11/30/22 was not accurate “4GB/s/bank with 512 IO @500Mbps”
  - **Peak bandwidth with open page:**
    Assuming continuous burst from a bank, this peak bandwidth is 512*500Mbps = **32GB/s**
  - **Effective bandwidth with closed page:**
    Assuming each access only reads 1 word of 512bits, bandwidth is
    512/tRC = 512/25ns = **2.56GB/s**
    To achieve 4GB/s, tRC of 16ns is needed.
- Bank size vs bandwidth
  - Using the proposed 2.5GB/s as a baseline, assuming tRC remains at 25ns.
  - Variation is bandwidth is achieved by varying the bank size in MB.
  - Bank physical size is normalized as mm^2/GB
  - Bank effective bandwidth is normalized as GB/s per GB
- Projecting from 25nm to 1-alpha
  - 1-alpha cell is ~0.55x in linear dimension. Overall array ~0.57x in linear dimension.
  - Assuming MAT cannot be customized, but array can be – need to be confirmed with foundry.
    - This means bank size under 1MB will not have 9 independent WLs, which is non-ideal for ECC.
  - HB pitch has to scale with linear dimensions.

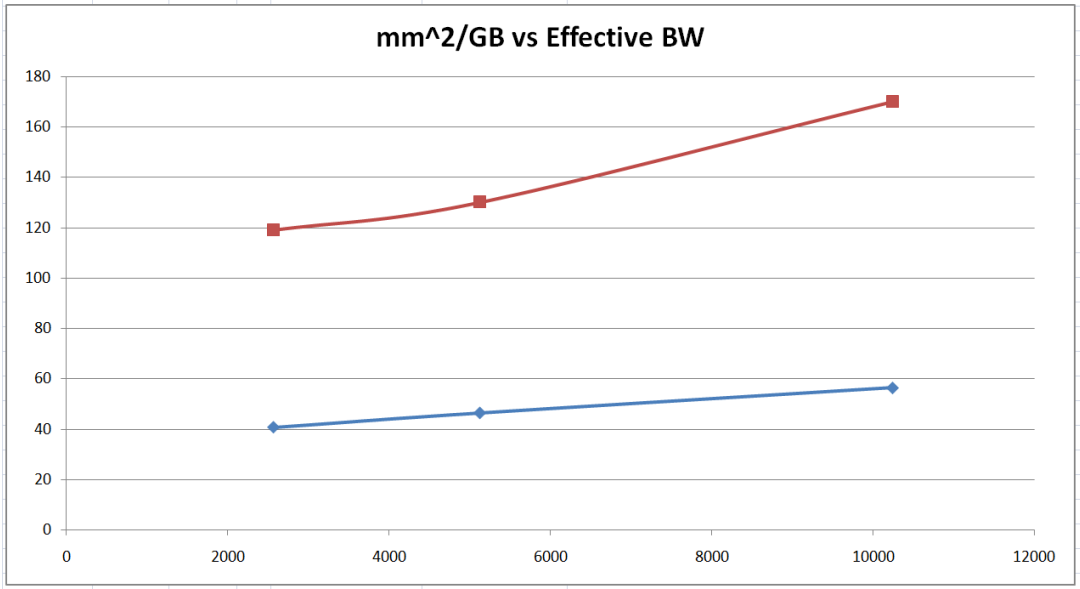
Die size vs Bandwidth

25nm

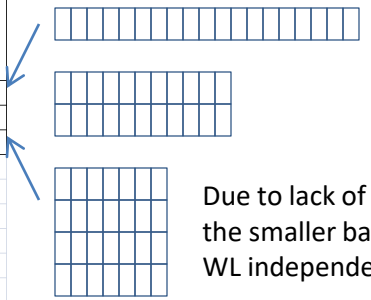
Bank mm2	Bank MB	Bank IO	tRC	Normalized Eff. BW/GB (GB/s)	Normalized bank size mm2/GB (10mm^2 added for fullchip overhead)	Notes
0.109	1.024	512	25	**2560**	**119**	HB pitch not limited; 9 independent WLs; 512x512 MAT
0.06	0.512	512	25	**5120**	**130**	HB at 2.5/5um X/Y pitch; 9 independent WLs; 512x512 MAT
0.04	0.256	512	25	**10240**	**170**	HB at 2.5/2.5um X/Y pitch; 9 independent WLs; 512x256 MAT

1-alpha

Bank mm2	Bank MB	Bank IO	tRC	Normalized Eff. BW/GB (GB/s)	Normalized bank size mm2/GB (5mm^2 added for full chip overhead)	Notes
0.035739	1.024	512	25	**2560**	**41**	HB pitch not limited; 9 independent WLs; 1024x512 MAT
0.0207286	0.512	512	25	**5120**	**46**	HB pitch 1.4/2.8um X/Y; 5 independent WLs; 1024x512 MAT
0.012866	0.256	512	25	**10240**	**56**	HB pitch 1.4/1.4um X/Y; 3 independent WLs; 1025x512 MAT



Bank shape



Due to lack of MAT customization the smaller banks lose WL independence.

Die size vs Latency

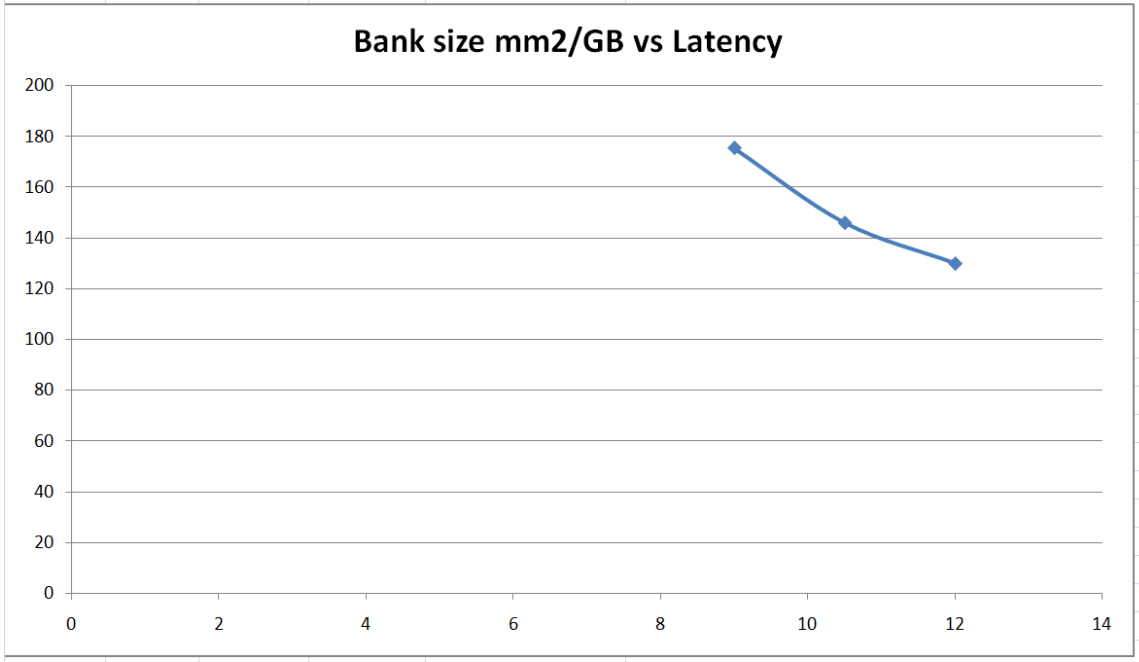
- Latency composition for proposed F25 case: 12ns = 2+8+2
  - This proposed case assume tRCD reduction of 3ns due to WL length reduction from 1024 to 512.
  - There may be opportunities to reduce to ~10ns
    - Experimental verification required. Tail bit behavior difficult to calculate.

- Further reduction in WL length 512 to 256 could reduce tRCD by 1.5ns
  - Further reduction has diminishing return
- Reduction in BL length from 512 to 256 could reduce tRCD by 1.5ns
- 1-alpha MAT customization is not possible. None of the above techniques are available to 1-alpha. We estimate 1-alpha latency to be in 15ns range.

Die size vs Latency

25nm

MAT	Bank MB	Bank size mm2	**Latency (ns)**	**Bank size mm2/GB**	Notes
512x512	0.512	0.06	**12**	**130**	Base case, assuming 512x512 customized MAT
512x256	0.512	0.068	**10.5**	**146**	Assumes WL length reduction reduces tRCD by 1.5ns
256x256	0.512	0.082688	**9**	**176**	Assumes BL length reduction further reduces tRCD by 1.5ns



Wafer BOW experience

- APM limits our wafer bow to be **<200um**
  - Measured at DRAM fab, but calibrated to HB fab measurement value after accounting for systematic measurement error between DRAM fab and HB fab.
- Evidence of problems >300um
  - We have seen evidence of wafer breakage linked to wafer bow when bow exceeds 300um.