Year	**Design**	**Organization**	**Die Size**	λ	λ^2 **area**	**cycle**	$\frac{\text{ALU bit ops}}{\lambda^2\text{s}}$
1985	[WDW+85]	1×16	15.5mm^2	1.0μ	16M	100 ns	10
1986	[vMWvW+86] [Gol87]	1×16	88.5mm^2	1.0μ	90M	125 ns	1.4
1987	[KNK+87]	1×16	11.5mm×12.9mm	0.65μ	350M	50 ns	0.9
1987	[CBBF87]	1×32	6mm×8.5mm	0.5μ	200M	60 ns	1.3
1989	[PML+89]	1×16	71mm^2	0.6μ	200M	100 ns	0.8
1992	[SKYH92]	1×16	9.5mm×10.5mm	0.6μ	275M	50 ns	1.2
1993	[USO+93]	1×16	9.3mm×9.1mm	0.4μ	530M	93 ns	0.33
1995	[NHK95]	1×32	8.5mm×8.5mm	0.25μ	1.2G	10 ns	2.8
1996	[KOT+96]	2×16	10mm×9.7mm	0.25μ	1.6G	25 ns	0.8

$$\text{ALU Bit Ops}/\lambda^2\text{s} = \frac{W_d \times N_{iALU}}{\frac{Area}{\lambda^2} \times t_{cycle}}$$

Sizes above include MACs, which generally amount to 5-10% of the total DSP die area.

Table 4.7: Survey of DSP Capacity

Year	**Design**	$F_{density}$	$I_{density}$	$D_{desnsity}$
1985	[WDW+85]	10	0	1.5×10^{-5}
1986	[vMWvW+86]	1.4	0	4.9×10^{-5}
1987	[KNK+87]	0.9	2.9×10^{-6}	9.3×10^{-5}
1987	[CBBF87]	1.3	1.0×10^{-5}	4.1×10^{-5}
1989	[PML+89]	0.8	1.6×10^{-5}	4.1×10^{-5}
1992	[SKYH92]	1.2	1.4×10^{-5}	3.0×10^{-5}
1993	[USO+93]	0.33	1.5×10^{-5}	1.5×10^{-4}
1995	[NHK95]	2.8	8.9×10^{-7}	2.9×10^{-5}
1996	[KOT+96]	0.8	?	?

Table 4.8: DSP Capacity Summary

4.4 Memories

Most general-purpose devices use memories to store instructions and data. A memory can also be used directly to implement complex computational functions. For complicated functions, a memory lookup can often provide high, programmable computational capacity. For just a few examples see [Has87, HT95, RS92, Lev77].

Model We characterize a memory by:

- depth – either in terms of address bits (a) or total number of memory words (2^a)
- width – w, the number of bits read out for each a bits of address put into the memory

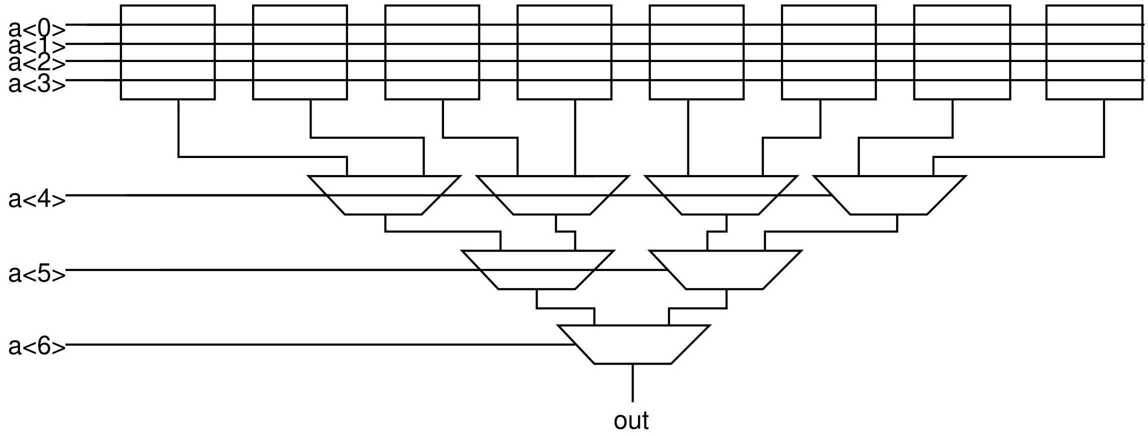

Figure 4.4: Gate Implementation of any Function Computed by 7-input Lookup Table

- t_{cycle} – the minimum time required between successive read operations

Capacity Provided The capacity provided by a memory is highly dependent on the inherent complexity of the logic function being computed. The lower bound is trivially zero since we could program the identify function into a memory.

We can use a counting argument to determine how complicated the functions can get. We start by observing that an a-input, one-output lookup table can implement 2^{2^a} different functions. We then consider how many gates it requires to implement any of these functions. Each gate can be any function of four inputs, so each gate can implement 2^{2^4} functions. A collection of n gates can thus implement at most $\left(2^{2^4}\right)^n$ functions (less due to overcounting). In order to implement any of the 2^{2^a} functions provided by the table, we need at least:

$$\begin{aligned} 2^{2^a} &\leq \left(2^{2^4}\right)^n = 2^{\left(n \cdot 2^4\right)} \\ 2^a &\leq n \cdot 2^4 \\ 2^{(a-4)} &\leq n \end{aligned} \tag{4.1}$$

Conversely, by construction, we can show that any function computed by the a input lookup table can be computed with $2^{(a-3)} - 1$ gate evaluations. As suggested in Figure 4.4, we can use $2^{(a-4)}$ gates to select the correct functional value based on the low four bits of the address. We then build a binary mux reduction tree to select the final output based on the remaining address bits. This tree requires $2^{(a-4)} - 1$ muxes. Together, the $2^{(a-3)} - 1$ gates compute any function computable by the a-input lookup table.

An a input by one output table lookup can thus provide between $2^{(a-4)} + 1$ and $2^{(a-3)} - 1$ gate evaluations per cycle for the most complicated a input functions. Since the bounds are essentially a factor of two apart, we can approximate the peak as $2^{(a-4)}$ gate evaluations per cycle. If the table is

w bits wide, the table provides at most w times as many gate evaluations. Putting all this together, we get:

$$C_{memory\_peak} \approx \frac{w \cdot 2^{(a-4)}}{\text{area} \cdot t_{cycle}} \tag{4.2}$$

Tables 4.9, 4.10 and 4.11 reviews memory implementations, showing the peak functional density for each memory array. For the most complex functions, memories provide the highest capacity of any general-purpose architecture. For less complex operations, however, memories are inefficient, yielding very little of their potential capacity.

For example, an 8-bit add operation with carry output requires 16 gate evaluations. Performed in a $2^{16} \times 9$ memory, such as a 9-bit version of the 64K$\times$18 memory from [SMK+94], this provides only 2.4 gate-evaluations/λ^2s. The inefficiency of the memory-based adder increases with operand size since the number of gate evaluations in an n-bit add increases as $2n$ whereas the memory area increases as $2^{2n} \times (n+1)$.

For all the memories listed, the capacity is based on continuous cycles of random access. In particular, nibble, fast page, or synchronous access in DRAMs is not exploited. For example, [TNK+94] achieves 13,500 gate-evaluations/λ^2s on random access. In sequential access mode, the part can output 18 bits every 8ns. For large sequential access, this means an effective cycle time of 8ns instead of the 48ns quoted – a factor of six improvement in cycle time and capacity. Used in this mode, the peak performance is 81,000 gate-evaluations/λ^2s.

It is also worth noting that, unlike processors, the capacity of memories has increased over time. This is likely due to:

- increased specialization of the fabrication processes to memories – especially the introduction of three-dimensional structures in DRAMs, local interconnect, and high poly film resistors for SRAMs
- increased pipelining of memory access

Modern processors actually dedicate a significant portion of their area to memory. Table 4.12 summarizes the peak capacity the processor can extract by using table lookups in its D-cache. The area used in calculating this capacity is the entire processor for the processors listed in Table 4.2. This peak capacity can be thought of as the peak capacity one could extract from each load operation when using the on-chip D-cache for table lookup operations.

Year	Design	Organization	Size	λ	λ^2 area	cycle	$\frac{\text{gate-evals}}{\lambda^2 s}$
1984	[BDN84]	16K$\times$1	16.3mm^2	1.0μ	16M	35 ns	1800
1984	[SMI$^+$84]	32K$\times$8	6.7mm$\times$8.9mm	0.6μ	164M	46 ns	2200
1984	[YKK$^+$84]	64K$\times$1	4.7mm$\times$6.6mm	0.75μ	55M	25 ns	3000
1984	[SCLB84]	4k$\times$16	32.6mm^2	0.75μ	58M	30 ns	2400
1984	[MSM$^+$84]	8K$\times$8	6.0mm$\times$6.8mm	0.6μ	113M	28 ns	1300
1984	[OKH$^+$84]	2K$\times$8	2.7mm$\times$3.5mm	0.75μ	17M	16 ns	3800
1984	[CH84]	4K$\times$4	11mm^2	1.05μ	10M	18 ns	5700
1984	[MMS$^+$84]	64K$\times$1	3.2mm$\times$6.0mm	0.65μ	45M	20 ns	4600
1985	[YTN$^+$85]	32K$\times$8	5.0mm$\times$9.2mm	0.65μ	110M	45 ns	3400
1985	[SAI$^+$85]	32K$\times$8	49.6mm^2	0.65μ	120M	45 ns	3100
1985	[SGS$^+$85]	8K$\times$8	31.8mm^2	0.75μ	57M	35 ns	2100
1985	[KEK$^+$85]	32K$\times$8	40.7mm^2	0.6μ	110M	55 ns	2600
1986	[CC86]	8K$\times$8	45mm^2	0.75μ	80M	35 ns	1500
1986	[KIK$^+$86]	256k$\times$1	4.5mm$\times$10.6mm	0.5μ	190M	25 ns	3500
1986	[FRV$^+$86]	64K$\times$1	3.4mm$\times$8.9mm	0.75μ	54M	13 ns	5900
1987	[WBS$^+$87]	32K$\times$8	6.8mm$\times$9mm	0.6μ	170M	21 ns	4600
1987	[KTO$^+$87]	128K$\times$8	8mm$\times$13.7mm	0.5μ	440M	35 ns	4300
1987	[WHS$^+$87]	128K$\times$8	5.5mm$\times$14.8mm	0.4μ	510M	34 ns	3800
1987	[MOT$^+$87]	128K$\times$8	6.9mm$\times$15.4mm	0.4μ	660M	25 ns	4000
1987	[GHS$^+$87]	32K$\times$8	9.5mm$\times$7.8mm	0.65μ	175M	40 ns	2300
1988	[SKI$^+$88]	128K$\times$8	7.6mm$\times$12.4mm	0.4μ	590M	44 ns	2500
1988	[WBEK$^+$88]	1M$\times$1	10.6mm$\times$8.5mm	0.35μ	730M	29 ns	3100
1988	[CDH$^+$88]	128K$\times$8	12.2mm$\times$7.7mm	0.35μ	770M	25 ns	3400
1988	[STT$^+$88]	256K$\times$4	7.5mm$\times$12mm	0.35μ	730M	18 ns	5000
1988	[SHU$^+$88]	256K$\times$4	6.2mm$\times$15.2mm	0.4μ	580M	15 ns	7500
1988	[KWA$^+$88]	1M$\times$1	5.5mm$\times$15.7mm	0.35μ	710M	14 ns	6600
1988	[ONN$^+$88]	32K$\times$8	4.4mm$\times$9.5mm	0.4μ	260M	7.5 ns	8400
1989	[VPP$^+$89]	256K$\times$1	3.9mm$\times$9.5mm	0.35μ	300M	14 ns	3900
1989	[MMK$^+$89]	512K$\times$8	7.5mm$\times$17.4mm	0.25μ	2.1G	25 ns	5000
1989	[SIY$^+$89]	1M$\times$1	5.3mm$\times$10.3mm	0.25μ	870M	9 ns	8300
1990	[FPH$^+$90]	256K$\times$1	11.6mm$\times$3.7mm	0.5μ	170M	8 ns	12000
1990	[ASO$^+$90]	4M$\times$1	7.7mm$\times$18.6mm	0.28μ	1.9G	15 ns	9200
1990	[HKM$^+$90]	4M$\times$1	8.4mm$\times$18.0mm	0.3μ	1.7G	20 ns	7800
1990	[SIS$^+$90]	512K$\times$8	7.2mm$\times$16.9mm	0.25μ	1.9G	23 ns	5900
1990	[OHK$^+$90]	512K$\times$8	7.8mm$\times$17.4mm	0.25μ	2.2G	23 ns	5200
1991	[CCS$^+$91]	32K$\times$16	11.1mm$\times$10.1mm	0.4μ	700M	2 ns	23500
1992	[SSN$^+$92]	32K$\times$8	6mm$\times$9mm	0.4μ	340M	200 ns	240
1992	[GOK$^+$92]	2M$\times$8	18.3mm$\times$12.5mm	0.2μ	5.7G	12 ns	15300
1992	[MKS$^+$92]	4M$\times$4	10.4mm$\times$21.5mm	0.2μ	5.6G	15 ns	12500
1992	[SIU$^+$92]	256K$\times$4	4mm$\times$7.4mm	0.15μ	1.3G	7 ns	7200
1993	[SKS$^+$93]	4M$\times$4	9.7mm$\times$21.9mm	0.18μ	6.9G	9 ns	16800
1993	[SUT$^+$93]	4M$\times$4	10.4mm$\times$10.6mm	0.13μ	7.1G	20 ns	7400
1994	[IKM$^+$94]	4M$\times$4	10.3mm$\times$20.9mm	0.2μ	5.4G	30 ns	6500

Table 4.9: Survey of Peak Memory Logic Capacity (SRAM)

Year	Design	Organization	Size	λ	λ^2 area	cycle	$\frac{\text{gate-evals}}{\lambda^2 \text{s}}$
1984	[MKS+84]	256K×1	6.3mm×6.3mm	0.75μ	70M	140 ns	1700
1984	[BCH+84]	256K×1	50mm^2	1.0μ	50M	150 ns	2200
1984	[MKM+84]	256K×1	30.2mm^2	0.8μ	47M	116 ns	3000
1984	[KFO84]	256K×1	46.8mm^2	0.6 μ	130M	100 ns	1250
1984	[SNT+84]	128K×8	9.4mm×8.1mm	0.5μ	300M	120 ns	1800
1985	[KCE+85]	1M×1	5.5mm×10.5mm	0.5μ	230M	160 ns	1800
1985	[KFM+85]	1M×1	5mm×13mm	0.6μ	180M	260 ns	1400
1985	[SFO+85]	1M×1	5mm×12.5mm	0.6μ	170M	190 ns	2000
1985	[TJ85]	256K×4	6.0mm×11.4mm	0.6μ	190M	190 ns	1800
1986	[FSO+86]	1M×1	4.4mm×12.3mm	0.6μ	150M	190 ns	2300
1986	[TTS+86]	4M×1	6.2mm×16.0mm	0.4μ	620M	300 ns	1400
1986	[FOW+86]	4M×1	7.8mm×17.5mm	0.5μ	550M	200 ns	2200
1986	[KAI+86]	64K×4	3.1mm×6.9mm	0.6μ	59M	200 ns	1400
1986	[HOW+86]	1M×1	4.8mm×13.2mm	0.6μ	180M	260 ns	1400
1987	[MNA+87]	4M×1	4.9mm×14.9mm	0.4μ	450M	220 ns	2600
1987	[KSE+87]	4M×1	6.4mm×17.4mm	0.4μ	690M	230 ns	1600
1987	[OFW+87]	4M×1	6.9mm×16.1mm	0.45μ	550M	220 ns	2200
1987	[MYM+87]	256K×4	4.7mm×13.8mm	0.5μ	260M	220 ns	1100
1988	[YKMI88]	256K×16	7.5mm×12.7mm	0.4μ	600M	200 ns	2200
1988	[LCwH+88]	128K×4	8.1mm×9.6mm	0.5μ	310M	36 ns	2900
1988	[ANH+88]	16M	8.2mm×17.3mm	0.3μ	1.6G	180 ns	3700
1988	[IYK+88]	16M×1	5.4mm×17.4mm	0.25μ	1.5G	120 ns	5800
1989	[WOI+89]	4M×4	17.5mm×12mm	0.35μ	1.7G	190 ns	3200
1989	[FOS+89]	16M×1	7.9mm×17.4mm	0.3μ	1.5G	150 ns	4600
1989	[CTK+89]	16M×1	8.0mm×16mm	0.28μ	1.7G	150 ns	4100
1989	[AFM+89]	16M×1	7.7mm×17.5mm	0.25μ	2.2G	120 ns	4100
1989	[CKC+89]	16M×1	8.5mm×18.4mm	0.3μ	1.7G	190 ns	3200
1989	[LBK+89]	256K×4	6.8mm×12.3mm	0.5μ	335M	56 ns	3500
1990	[TTK+90]	16M×1	8.2mm×15.9mm	0.28μ	1.7G	150 ns	4100
1990	[KDK+90]	4M×1	5.6mm×15.2mm	0.35μ	700M	120 ns	3100
1990	[KSB+90]	16M×1	7.8mm×18.1mm	0.25μ	2.3G	150 ns	3100
1991	[NTT+91]	16M×4	9.7mm×20.3mm	0.15μ	8.8G	180 ns	2700
1991	[MMM+91]	64M×1	12.5mm×18.7mm	0.2μ	5.8G	120 ns	6000
1991	[TTU+91]	64M×1	19.9mm×11.3mm	0.2μ	5.6G	120 ns	6200
1991	[OTW+91]	64M×1	9.2mm×19.1mm	0.2μ	4.4G	90 ns	11000
1991	[YNH+91]	4M×16	234mm^2	0.2μ	5.6G	95 ns	7500
1991	[NNO+91]	4M×1	4.8mm×11.1mm	0.3μ	592M	60 ns	7400
1992	[HAH+92]	16M×1	8mm×16.6mm	0.3μ	1.5M	120 ns	6000
1992	[KDK+92]	512K×8	13.2mm×6.4mm	0.4μ	525M	60 ns	8300
1993	[STN+93]	16M×16	13.6mm×24.5mm	0.13μ	21G	60 ns	13100
1993	[KHK+93]	64M×4	14.4mm×33.2mm	0.13μ	30.6G	100 ns	5500
1994	[TNK+94]	1M×18	17.1mm×6.6mm	0.25μ	1.8G	48 ns	13500
1994	[AOT+94]	32M×8	13.3×22.8mm	0.13μ	19.5G	90 ns	9600
1994	[TTT+94]	32M×8	13.2mm×25.9mm	0.13μ	21.9G	56 ns	13700
1995	[SMK+94]	32K×9	1.7mm×5.0mm	0.4μ	53M	50 ns	7000
1995	[SMK+94]	64K×18	2.1mm×4.9mm	0.25μ	165M	40 ns	11200

Table 4.10: Survey of Peak Memory Logic Capacity (DRAM)

Year	Design	Organization	Size	λ	λ^2 area	cycle	$\frac{\text{gate-evals}}{\lambda^2 s}$
Pseudostatic							
1984	[KSY+84]	32K×9	55mm^2	1μ	55M	125 ns	2700
1991	[SKK+91]	512K×8	6.5mm×14.2mm	0.4μ	580M	116 ns	3900
Virtually Static							
1986	[NSS+86]	128K×8	6mm×13.8mm	0.5μ	330M	150 ns	1300

Table 4.11: Survey of Peak Memory Logic Capacity (Hybrid)

Year	Design	Ref.	D-Cache $\frac{\text{gate-evaluations}}{\lambda^2 s}$
1984	RISC II	[SKPS84]	0
1984	MIPS	[RPJ+84]	0
1987	MIPS-X	[HHC+87]	0
1987	PA-RISC	[YFJ+87]	0
1990	PA-RISC	[TLB+90]	0
1990	SPARC	[MMN+90]	39
1992	SuperSparc	[ANAB+92]	250
1992	Alpha	[DWA+92]	610
1994	PA-RISC	[RDB+94]	0
1994	MIPS	[SYN+94]	150
1995	PowerPC	[BBB+95]	510
1995	UltraSparc	[CDd+95]	330
1995	SPARC V9	[SPA+95]	0
1995	Alpha	[BAB+95]	580
1996	MIPS	[KDS+96]	170
1996	PA-RISC	[LLNK96]	0
1996	ARM	[MWA+96]	1000
1996	Alpha	[GBB+96]	550

Table 4.12: Survey of Processor On-Chip Memory Capacity

4.5 Field-Programmable Gate Arrays (FPGAs)

Field-Programmable Gate Arrays (FPGAs) are composed of a collection of programmable gates embedded in a programmable interconnect. Programmable gates are often implemented using small lookup tables. The small lookup tables with programmable interconnect allow one to take advantage of the structure inherent in many computations to reduce the amount of memory and space required to implement a function versus the full memory arrays of the previous section. Ultimately, this allows FPGA space required for an application to scale with the complexity of the application rather than scaling exponentially in the manner of pure memories.

Model For pedagogical purposes, we consider an FPGA composed of:

- n, four-input lookup tables (4-LUTs) for gates with an optional flip-flop on the output of each LUT which can be used for pipelining or data storage
- "adequate" programmable interconnect to wire up functions using the 4-LUTs
- a minimum operating cycle time, t_{cycle}, which accounts for the time to travel through one LUT and its associated interconnect.

Year	Design	Organization	Size	λ	λ^2 area	cycle	$\frac{\text{gate-evals}}{\lambda^2\text{s}}$
1986	Xilinx 2K [CDF+86]	1 CLB (4-LUT)	$693\mu \times 715\mu$	1μ	500K	20 ns	100
1988	Xilinx 3K [Xil89, HDJ+88]	64 CLBs (2×4-LUT/CLB)	5mm×6mm (XC3020 die)	0.6μ	83M	13 ns	120
1991	UTFPGA [CSA+91]	3 4-LUTs	$900\mu\times800\mu$	0.6μ	2M	7 ns	210
1992	Xilinx 4K [Xil94b]	49 CLBs (2×4-LUT/CLB)	4.8mm×4.6mm (XC4005 Quadrant)	0.6μ	61M	7 ns	230
1994	LEGO [Seo94]	4 4-LUTs	$1240\mu\times1184\mu$	0.6μ	4M	4.1 ns	240
1995	DPGA [TEC+95]	16 4-LUTs	$1500\mu\times1750\mu$	0.5μ	10.5M	7 ns†	210
1995	Xilinx 5K [Xil91]	49 CLBs (4×4-LUTS/CLB)	3mm×3.3mm (XC5206 Quadrant)	0.3μ	110M	6 ns	290
1995	Altera Flex 8K [Alt95]	1008 LEs (4-LUT/LE)	8mm×10.5mm (81188A die)	0.3μ	930M	7.5 ns	144
1995	ORCA 2C [ATT95]	256 PLCs (4×4-LUT/PLC)	10mm×9.8mm (ATT2C10 die)	0.3μ	$1.1\text{G}\lambda^2$	7 ns	134

† No context switch – 10 ns cycle for context switch

$$\text{gate-evals}/\lambda^2\text{s} = \frac{N_{4LUT}}{\frac{Area}{\lambda^2} \times t_{cycle}}$$

Table 4.13: Survey of FPGA Capacity