

SCIENCES

Electronics Engineering, Field Director – Francis Balestra

Design Methodologies and Architecture, Subject Head – Ahmed Jerraya

Multi-Processor System-on-Chip 2

Applications

Coordinated by

Liliana Andrade
Frédéric Rousseau

Color Section

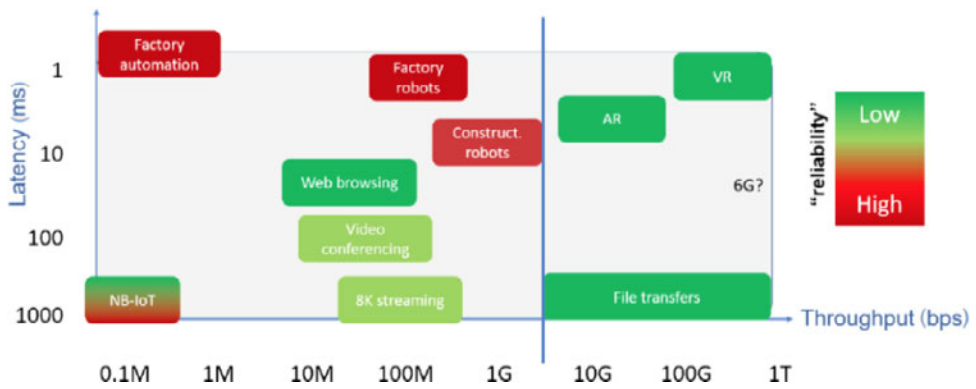


Figure 1.1. Application mapping on the rate–latency plane with regard to the reliability requirement (Fettweis et al. 2019)

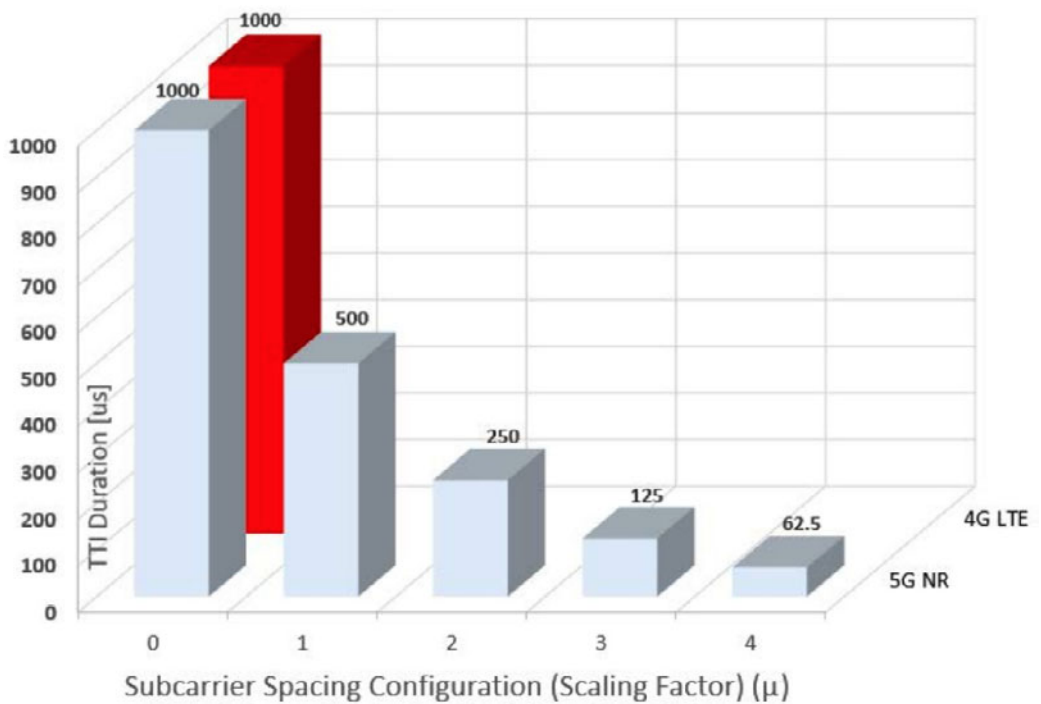


Figure 1.2. Comparing 14 OFDM symbols' TTI duration of 4G and 5G

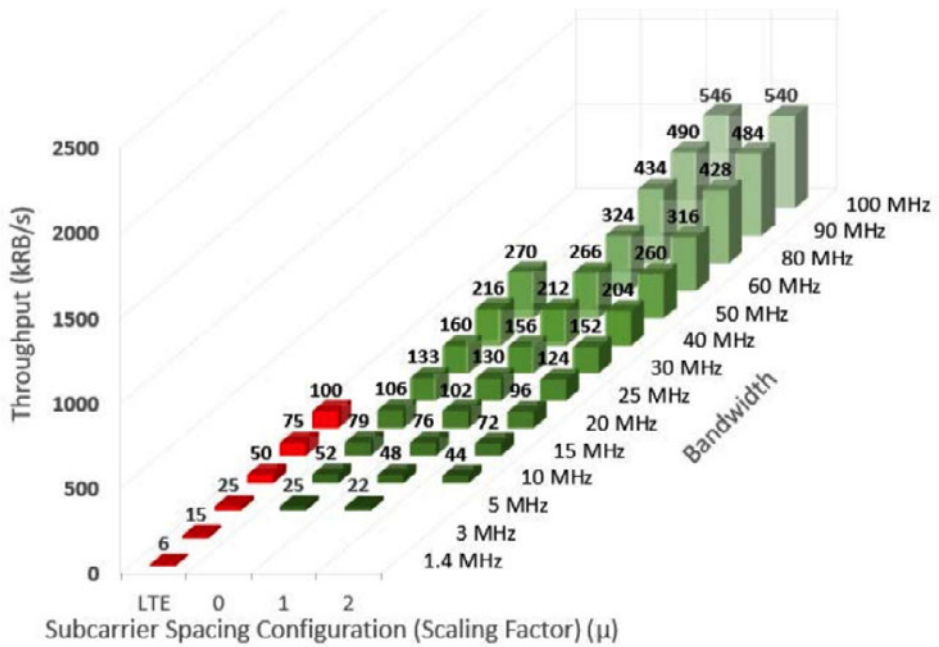


Figure 1.3. Processing load in kRB/s for 5G NR FR1 (Damjancevic et al. 2019)

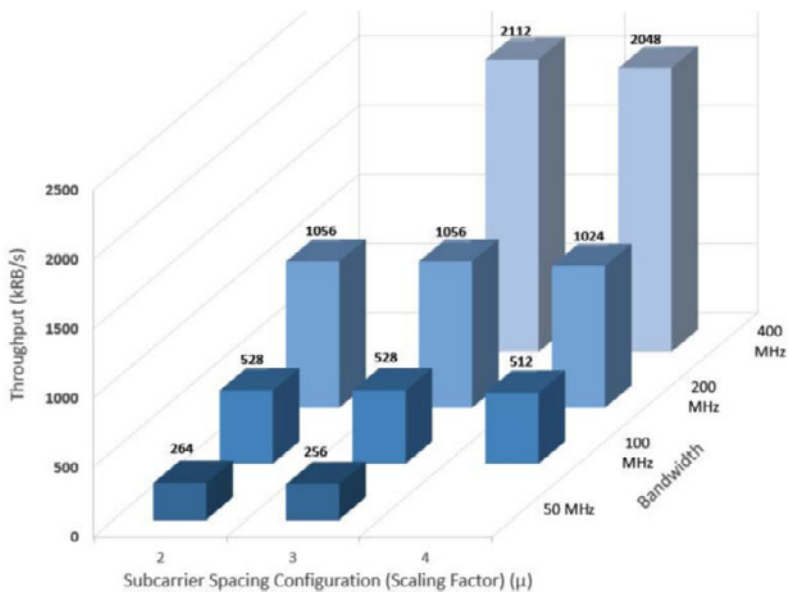


Figure 1.4. Processing load in kRB/s for 5G NR FR2

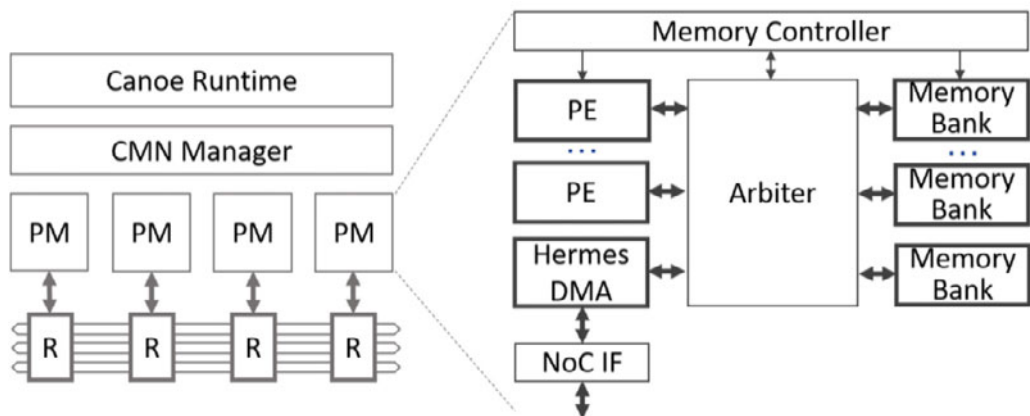


Figure 1.5. Tiled “Kachel” MPSoC with decentralized tightly coupled memories

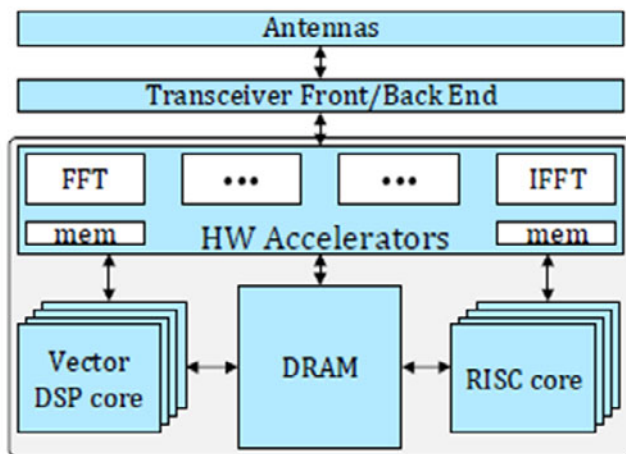


Figure 1.6. Heterogeneous MPSoC with a central shared memory architecture

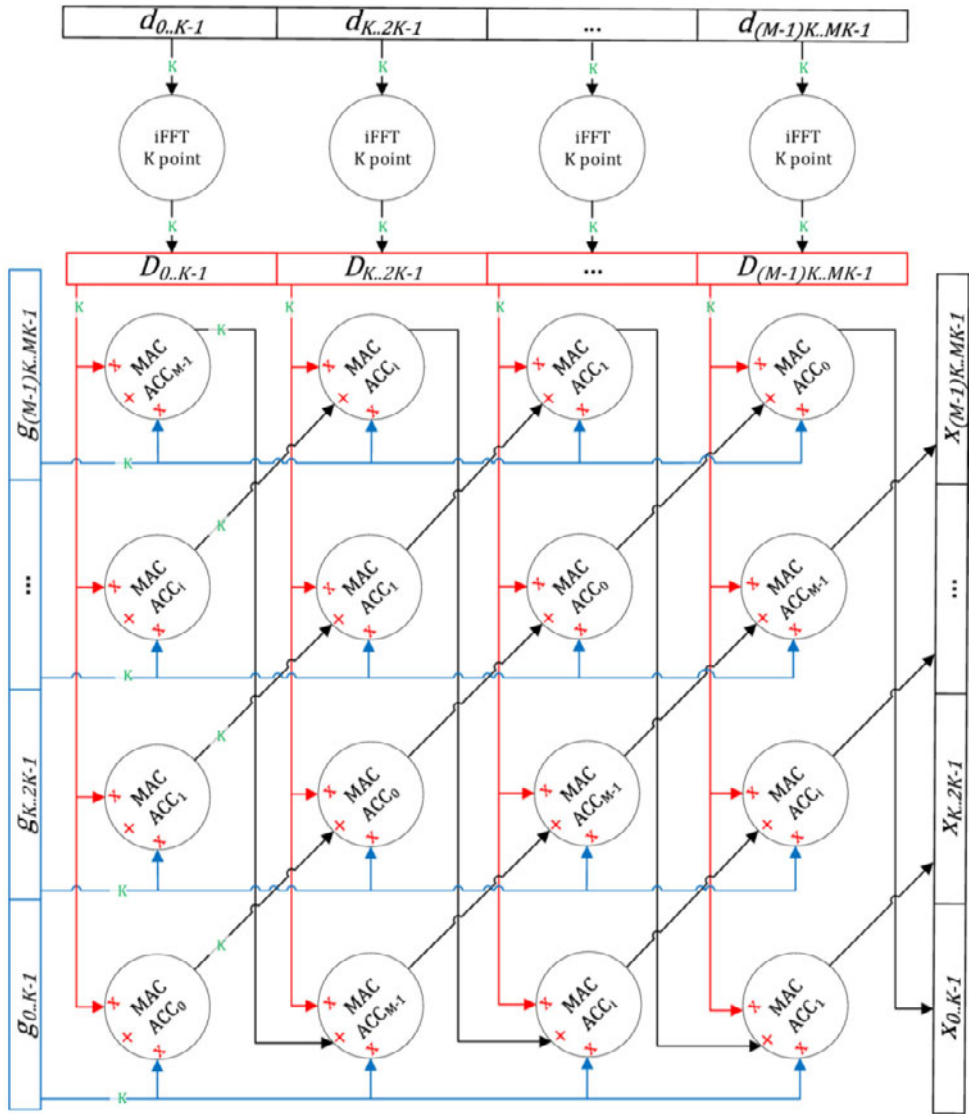


Figure 1.7. GFDM processing dataflow diagram

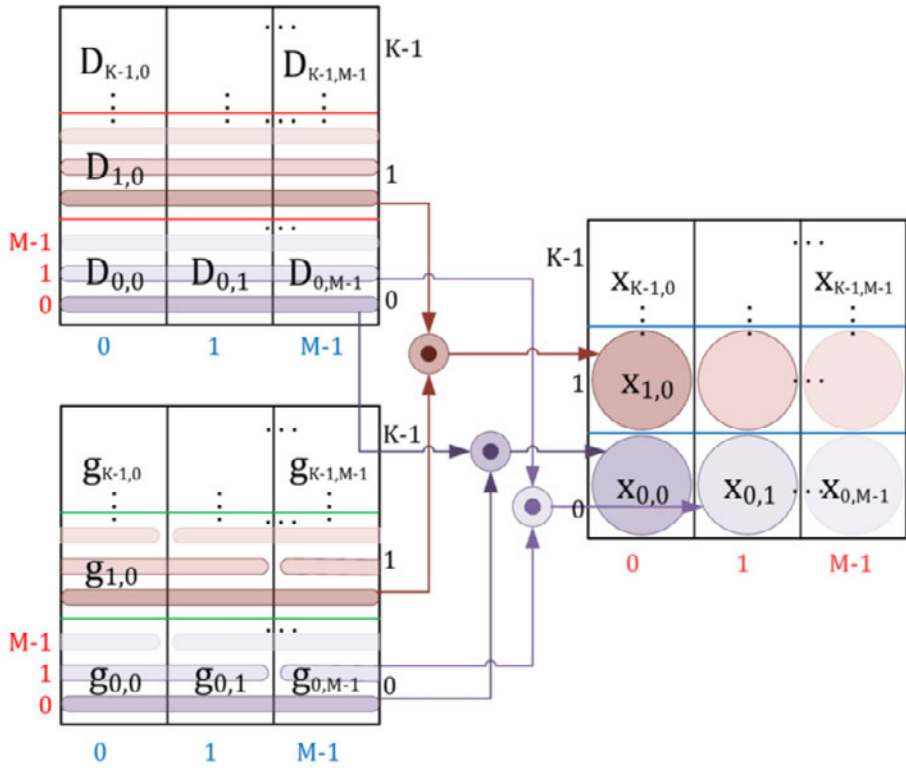


Figure 1.8. Visualization of time-domain GFDM filtering

```

input : IDFT output D and filter coefficients g
output: GFDM signal x

1 // Assume complex operations *,+=
2 for l ← M - 1 to 0 do
3   for i ← 0 to K - 1 do
4     ACC = 0;
5     for m ← 0 to l do
6       // split in 2 loops to avoid modulo operations
7       ACC += D[mK + i] * g[(M - 1 + m - l)K + i];
8     end
9     for m ← l + 1 to M - 1 do
10      ACC += D[mK + i] * g[(m - 1 - l)K + i];
11    end
12    x[(M - 1 - l)K + i] = ACC;
13  end
14 end

```

Figure 1.9. GFDM pseudo-code

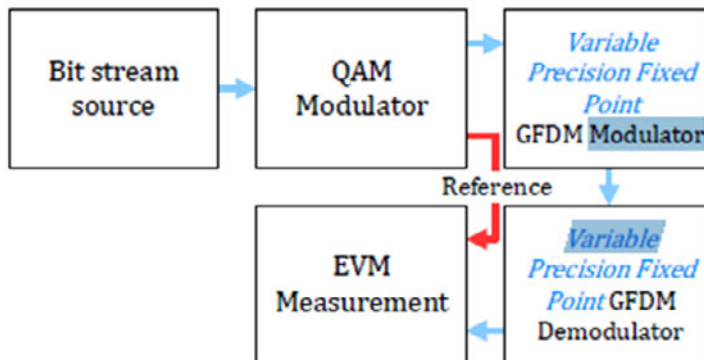


Figure 1.10. Precision test bed set up

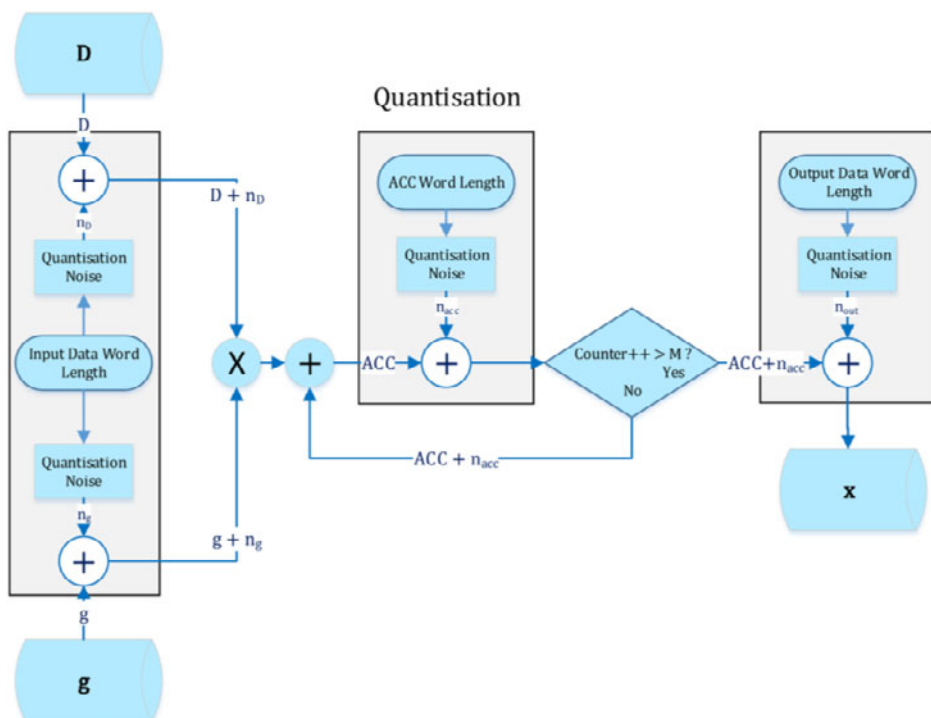


Figure 1.11. Varied precision quantization of GFDN

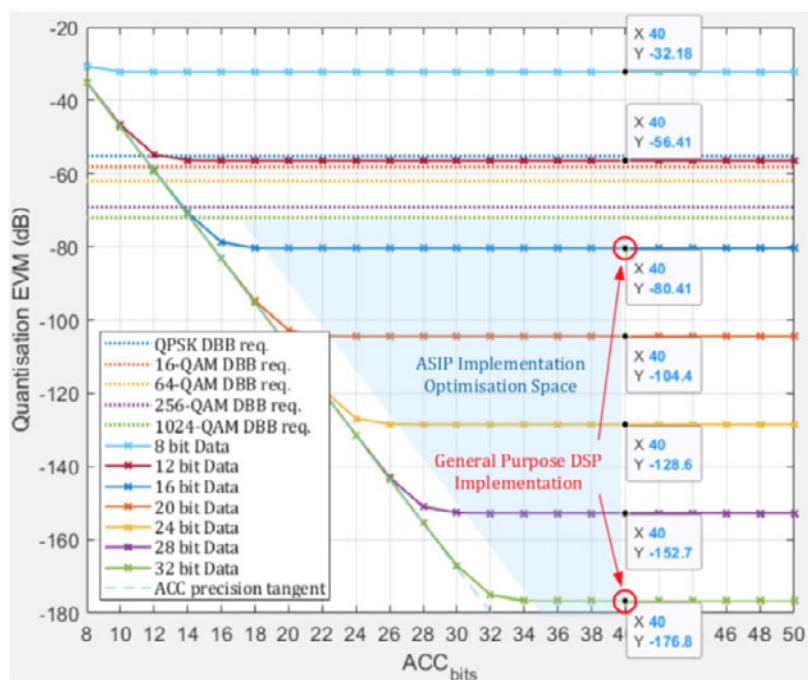


Figure 1.12. GFDN EVM for varied data and ACC complex bit-lengths compared to adjusted 3GPP EVM DBB requirements (3GPP 2018b, 2019a)

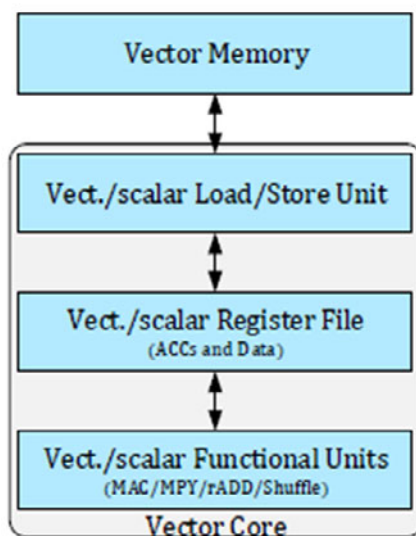


Figure 1.13. vDSP simplified HW block diagram

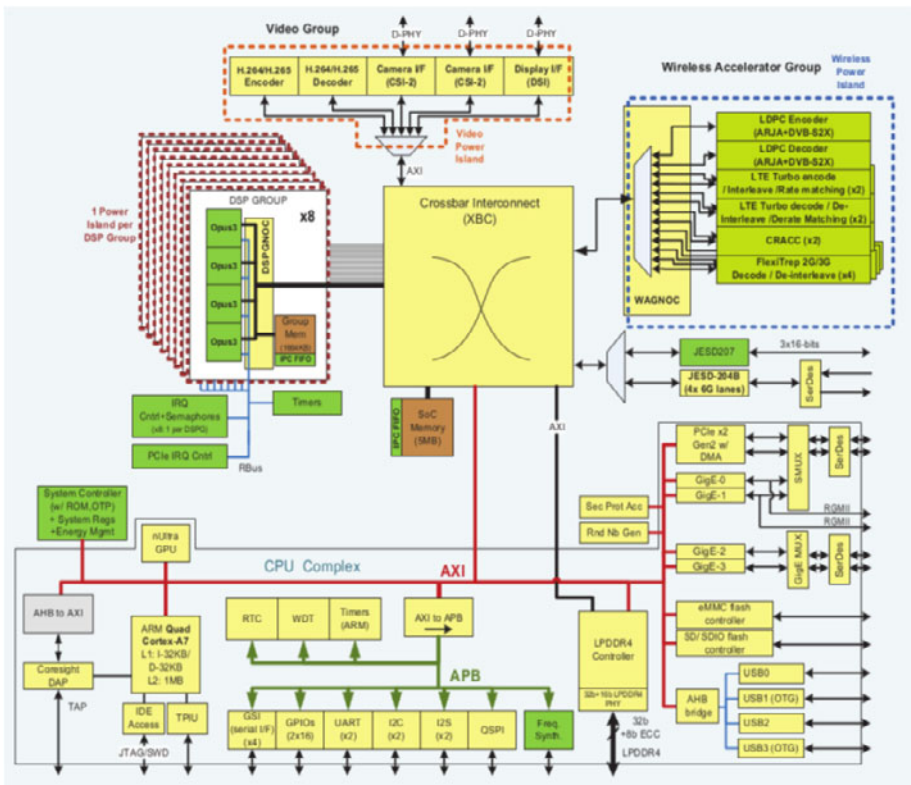


Figure 2.1. State-of-the-art commercial system-on-chip baseband architecture

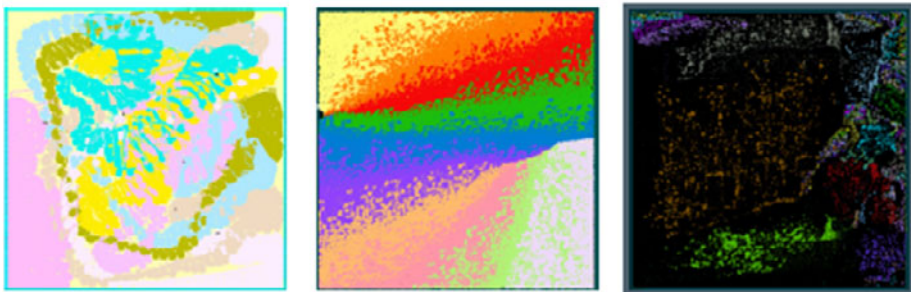


Figure 2.2. Left: 306Gbit/s turbo decoder. Middle: 288 Gbit/s LDPC decoder. Right: 764Gbit/s polar decoder

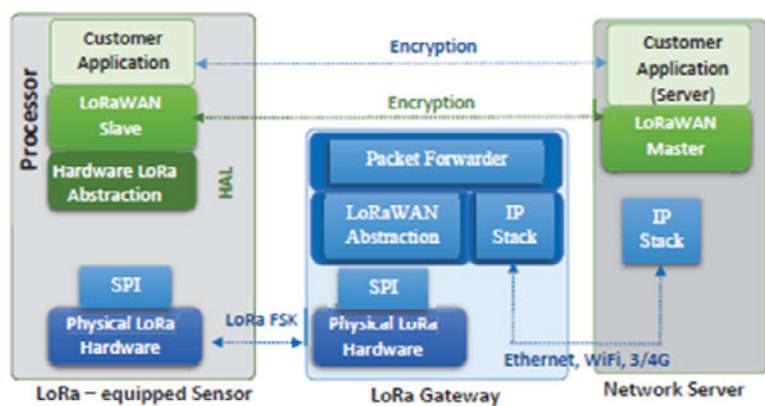


Figure 3.1. Security in LoRaWAN

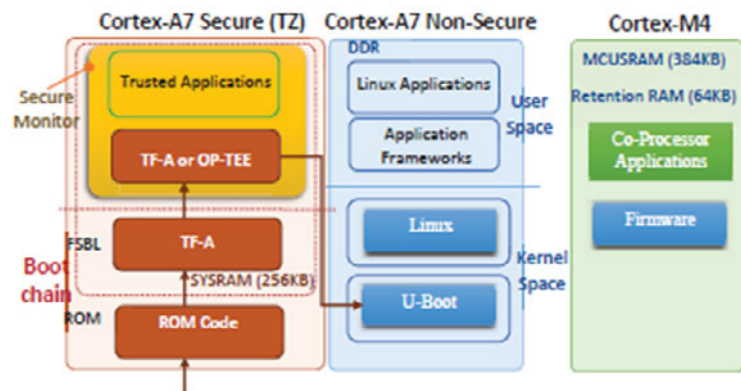


Figure 3.2. Boot process in an STM32MP1 device

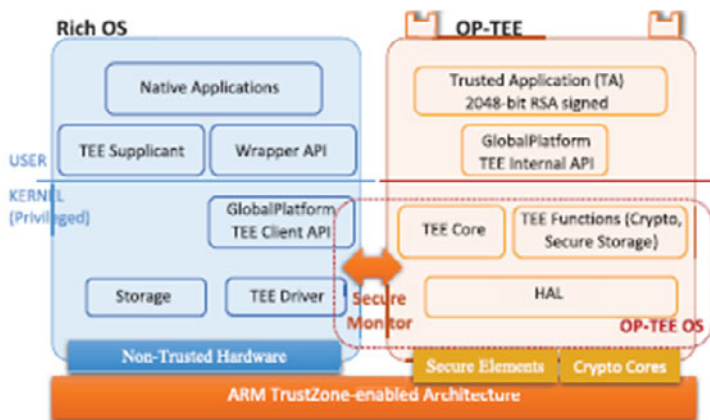


Figure 3.3. Execution environments in OP-TEE enabled organization based on ARM TrustZone architecture



Figure 3.4. LoraWAN gateway using an RAK831 RF with a GPS (top two shields), the STM32MP1 and a 3G modem (bottom shield) when a wired connection is not available

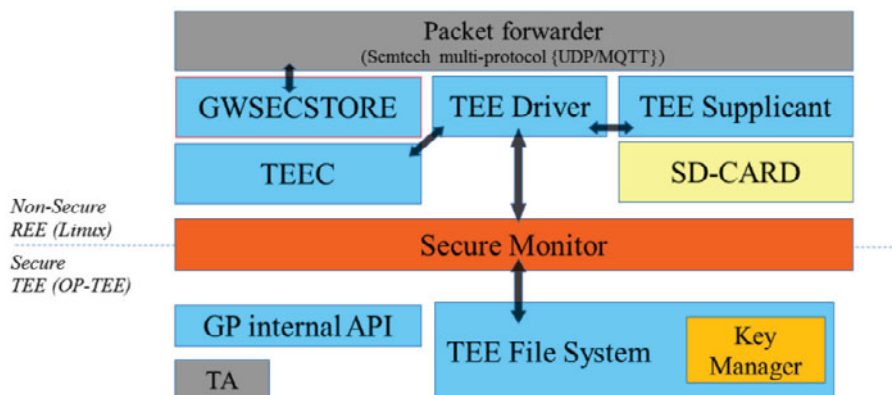


Figure 3.5. Execution of gateway packet forwarder in OP-TEE enabled organization based on the STM32MP1 platform

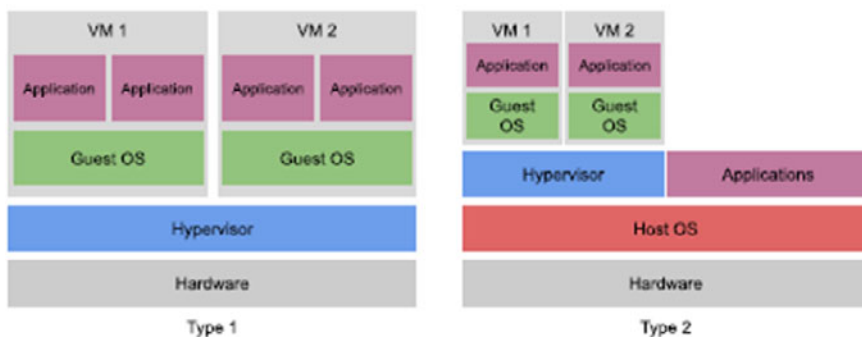
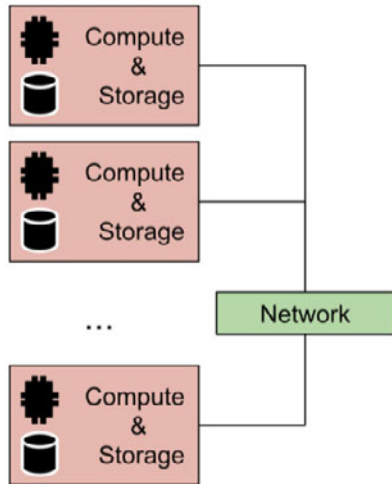


Figure 4.1. Hypervisor types

HyperConverged (HCI) Architecture



Disaggregated Architecture

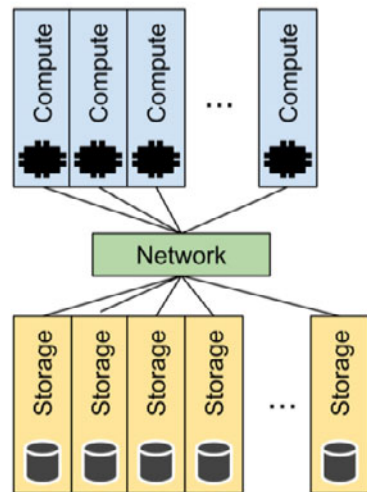


Figure 4.2. Hyperconverged versus disaggregated architectures

NexVisor I/O Architecture vs. Standard Xen Hypervisor

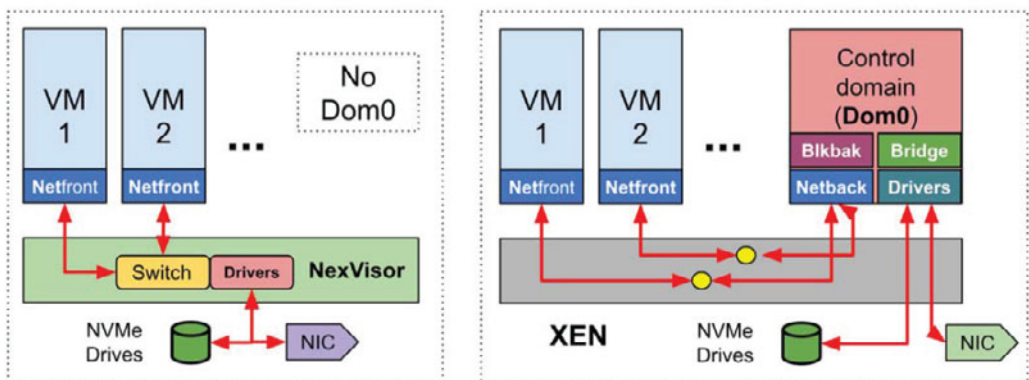


Figure 4.3. Comparison of the NexVisor I/O architecture to standard Xen

NexVisor Optimized I/O Architecture

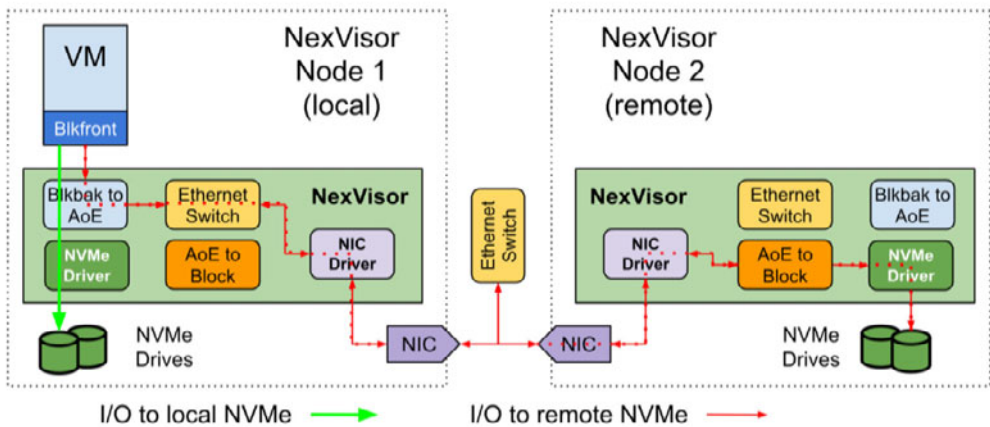


Figure 4.4. Optimized I/O datapath operations in the NexVisor for local and remote NVMe over the network

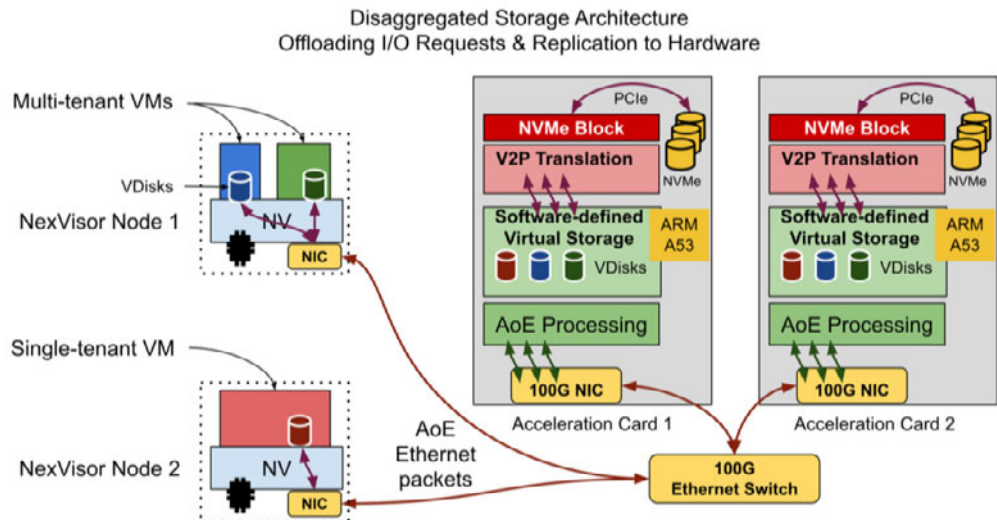


Figure 4.5. High-level view of disaggregated storage architecture, showing the I/O requests originating from the VMs through the NexVisors and through the ethernet network to the acceleration cards

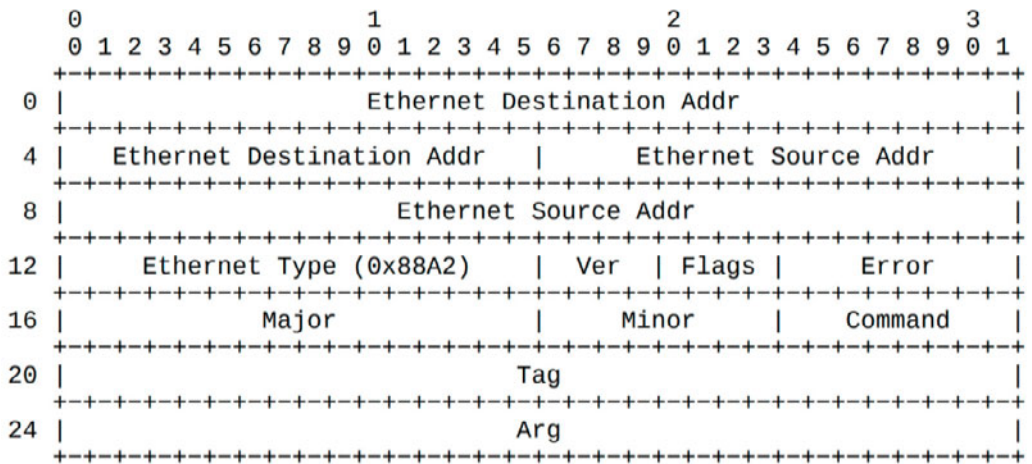


Figure 4.6. ATA over Ethernet (AoE) header format (Hopkins and Coile 2009)

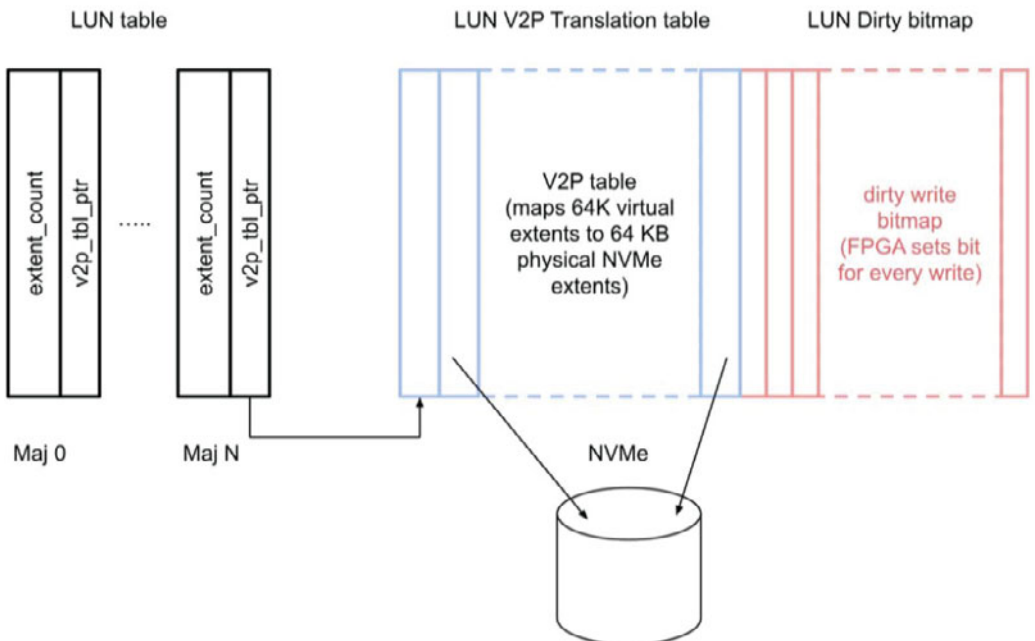


Figure 4.7. Storage virtualization data structures used by the accelerated datapath

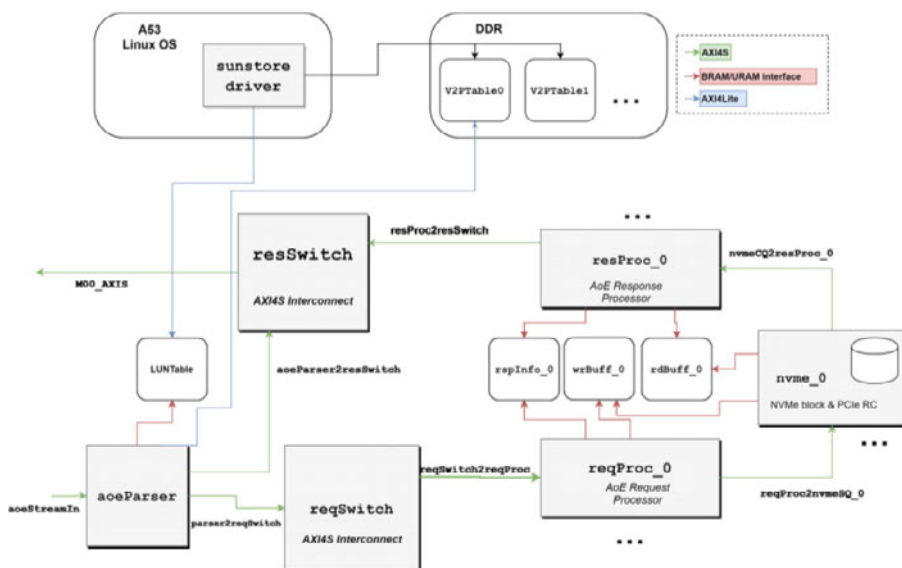


Figure 4.8. Hardware architecture for the disaggregated storage acceleration card

kIOPs: seqread on separate discarded disks (10Gbit/s link)

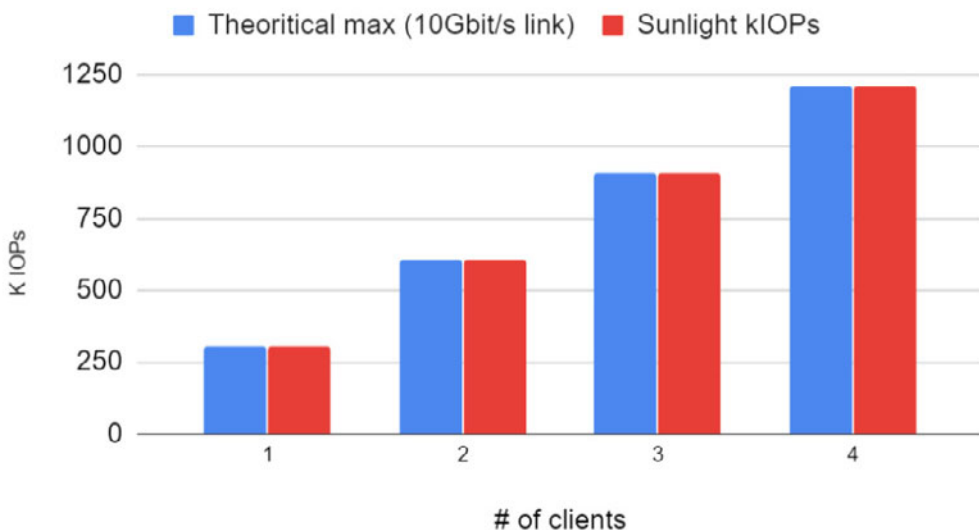


Figure 4.9. Thousands of sequential read I/Os per second, using fio on four VM clients, each reading on a separate NVMe drive

KIOPs: seqwrite on separate disks (10Gbit/s link)

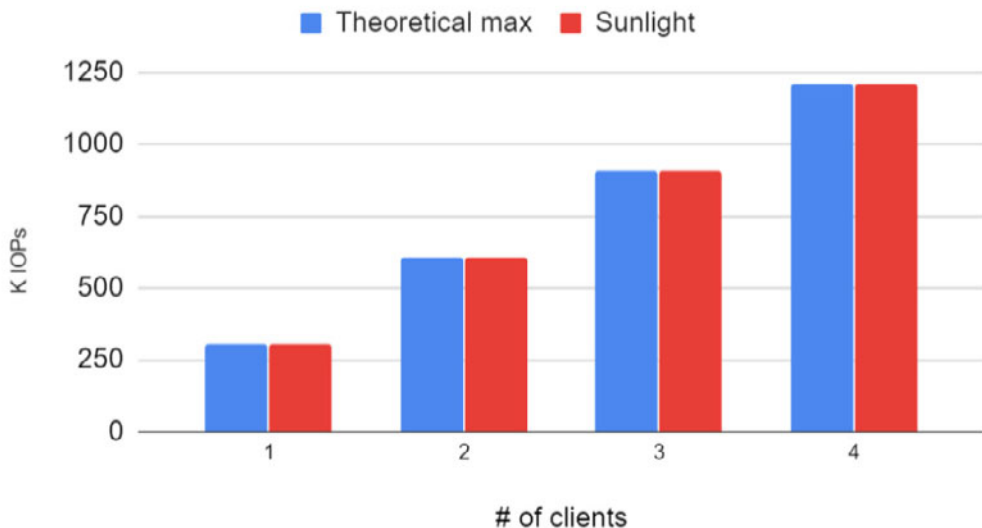


Figure 4.10. Thousands of sequential write I/Os per second, using *fiio* on four VM clients, each writing to a separate NVMe drive

KIOPs: seqread on same disk (10Gbit/s link)

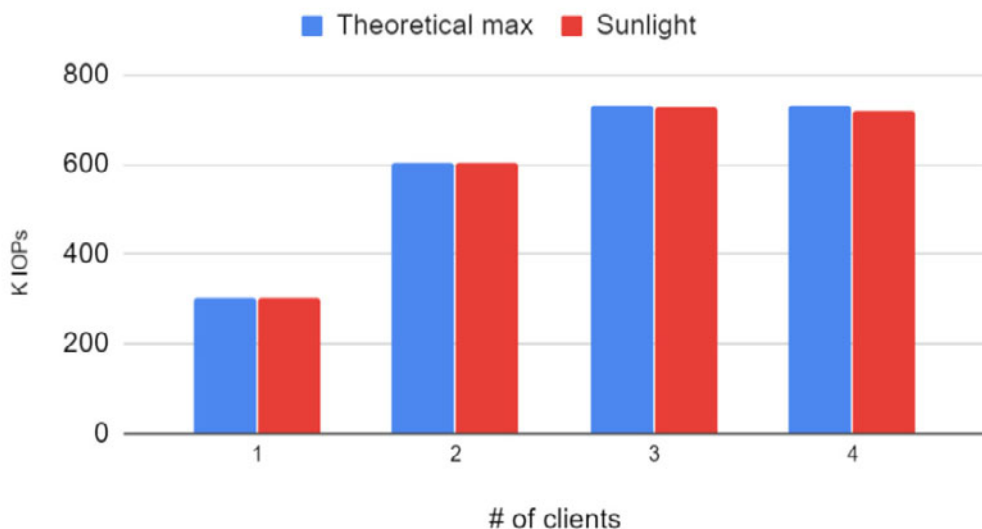


Figure 4.11. Thousands of sequential read I/Os per second, using *fiio* on four VM clients, all reading from the same NVMe drive

AoE Read throughput on separate discarded disks

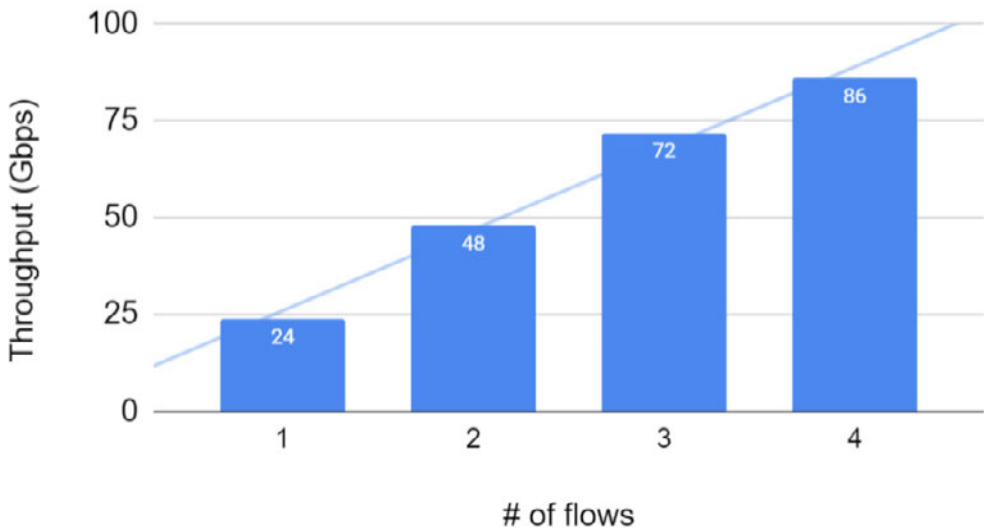


Figure 4.12. AoE read throughput scaling for one to four client flows

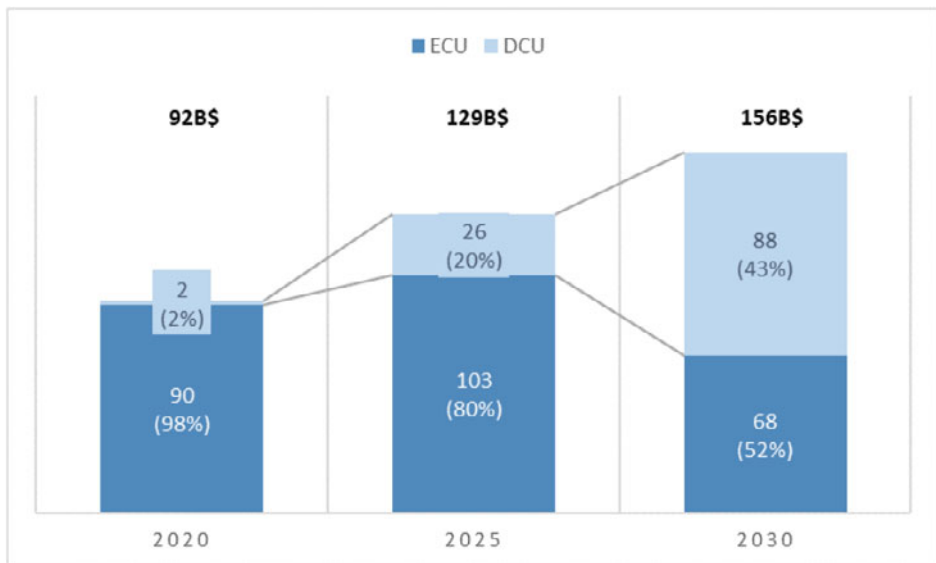


Figure 5.1. Overall ECU/DCU costs (B\$) evolution and breakdown between standard ECU and DCU from 2020 to 2030 (McKinsey and Company 2019)

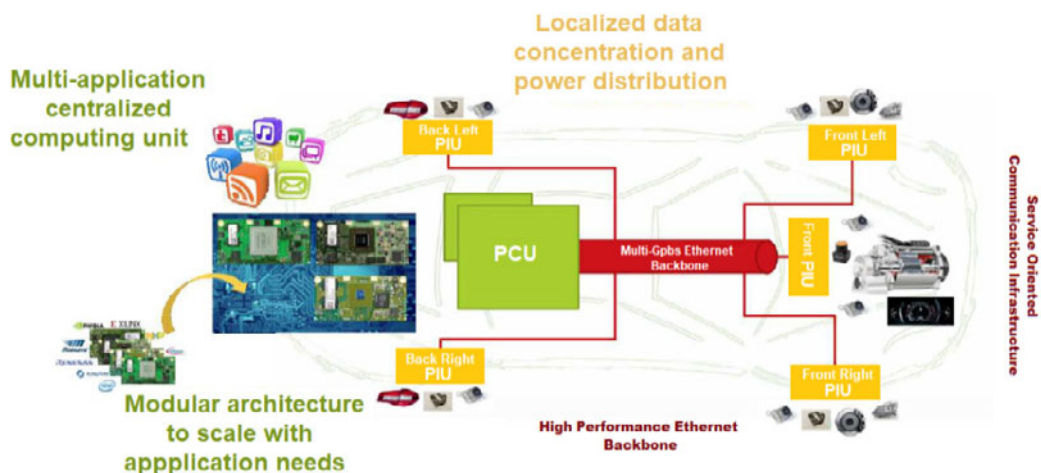


Figure 5.2. Overview of the FACE PCU and PIU infrastructure







Computing Platform	Bring-up	Peak Performances	Versatility	Criticality Management	Multi-Appli	Programmability
Infotainment 	-	+	+	++	+	++
ADAS/AD SoC 	-	+	+	+	+	+
DSP 	-	+	-	-	+	+
Data crunching 	++	++	--	--	-	++
FPGA platforms 	+	+	+	+	+	-
Many-core 	-	-	-	+	++	-

Figure 5.3. Synthetic result of the hardware benchmark

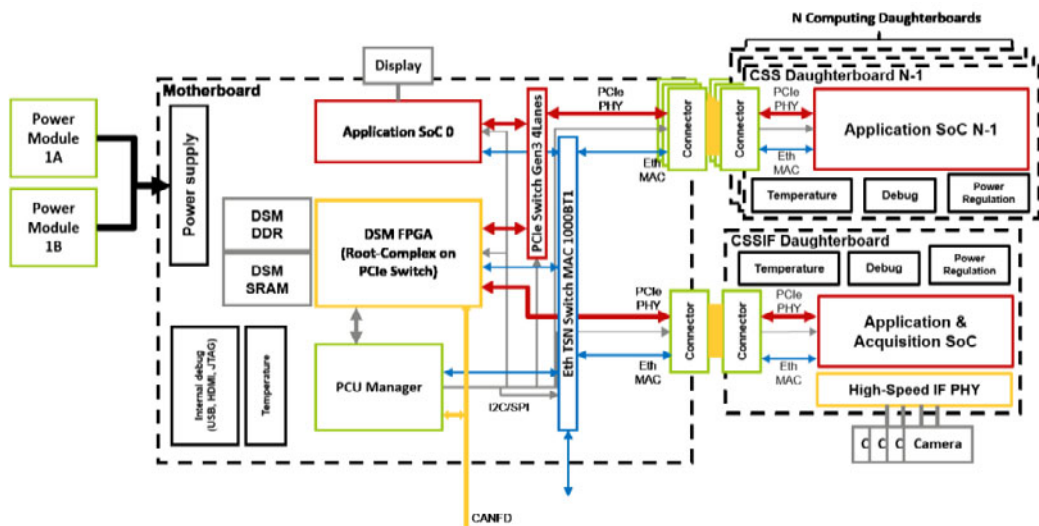


Figure 5.4. Overview of the FACE PCU structure

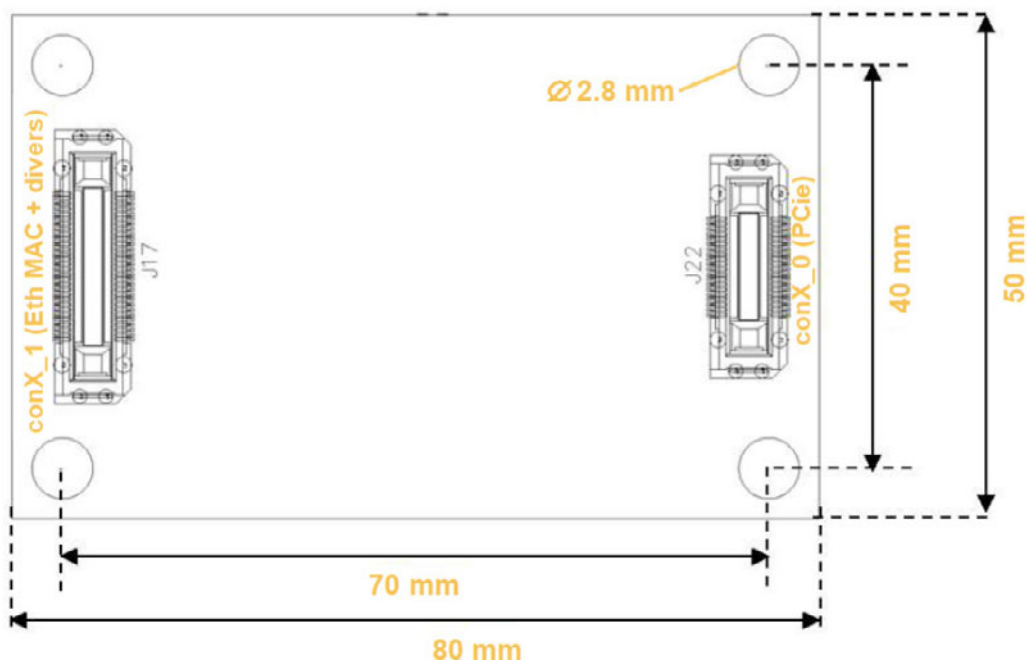


Figure 5.5. Daughterboard physical form factor

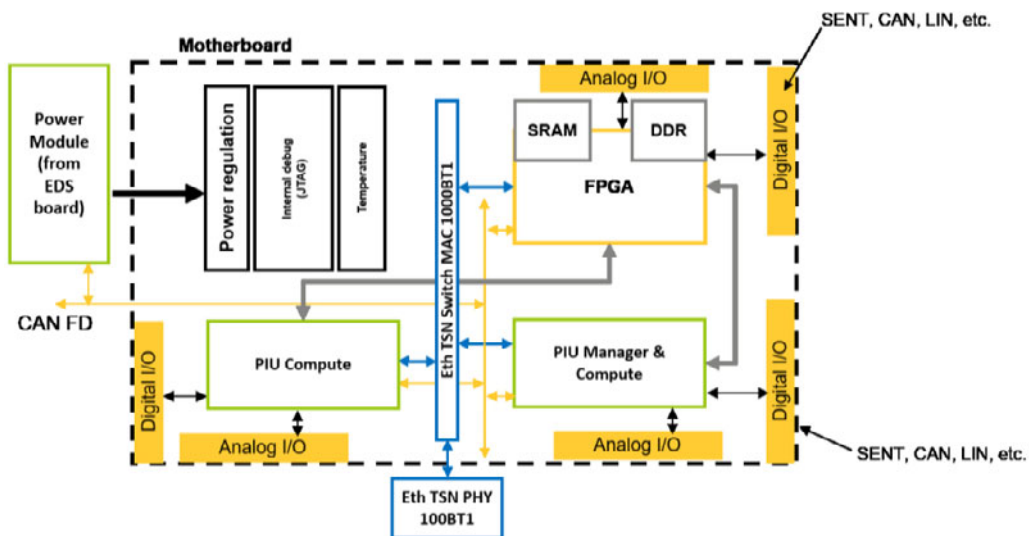


Figure 5.6. Overview of the FACE PIU structure

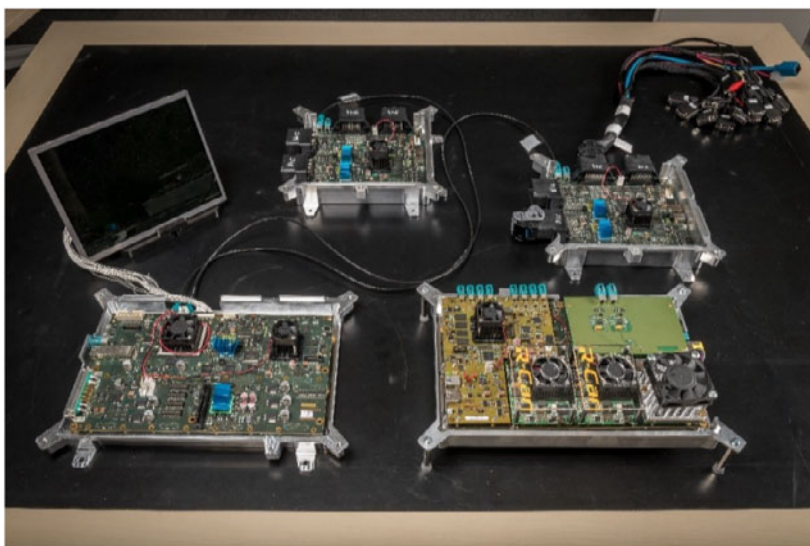


Figure 5.7. Example of hardware setup of the FACE platform. PCU front and back sides are on the bottom of the photo. PCU is connected through 100Base-T1 Ethernet TSN to two PIUs. Harness to connect heterogeneous sensors and actuators is in the right-top corner of the photo

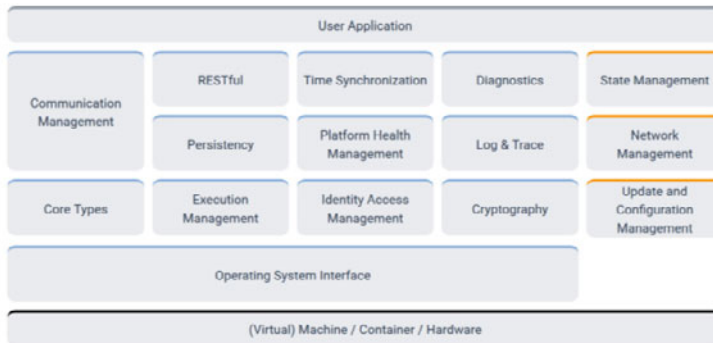


Figure 5.8. Illustration of AUTOSAR adaptive platform software architecture

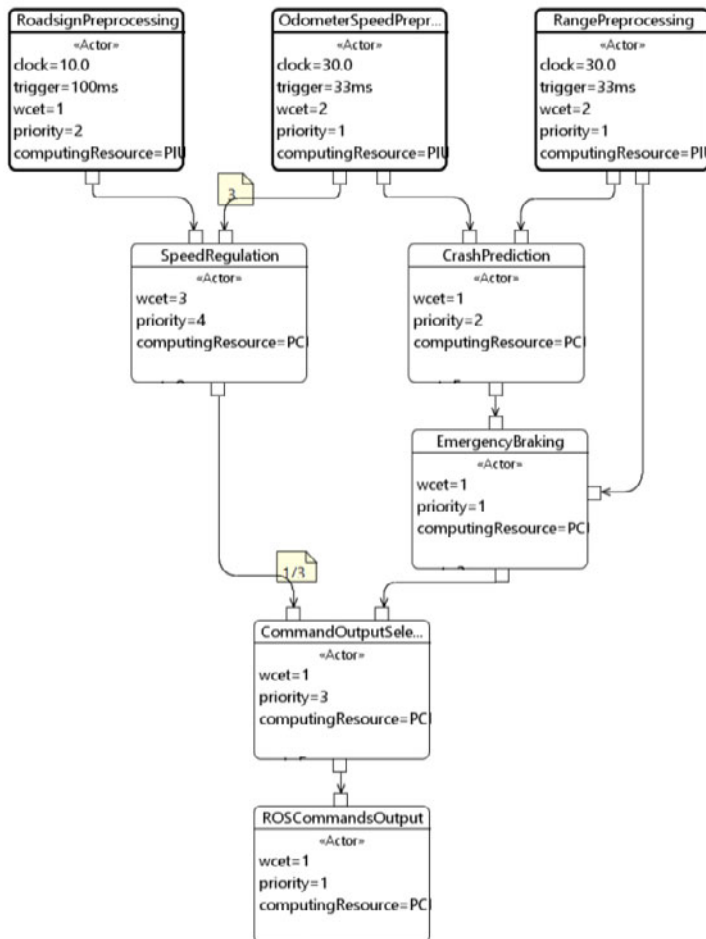


Figure 5.9. ADAS polygraph model

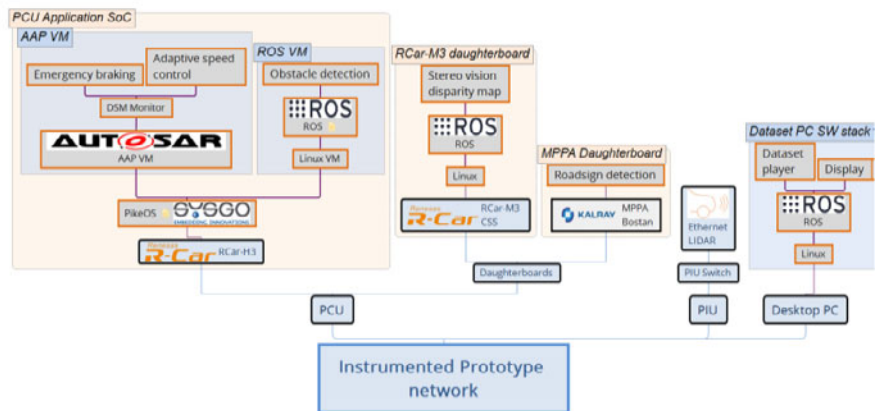


Figure 5.10. The FACE instrumented prototype setup

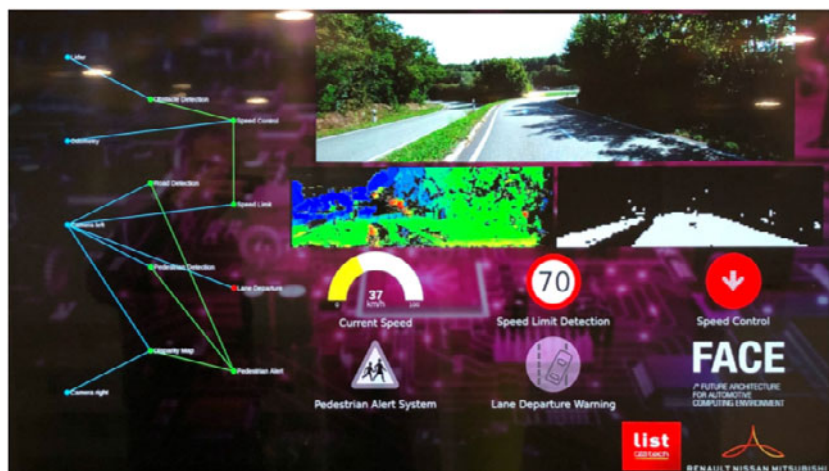


Figure 5.11. Use case interface

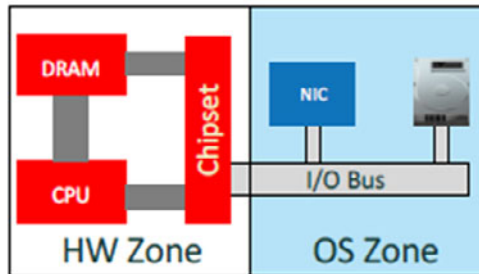


Figure 6.1. *The anatomy of a desktop in the 1980s*

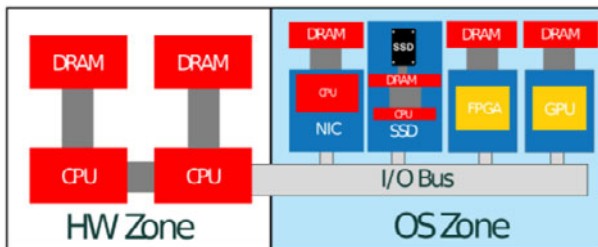


Figure 6.2. *The anatomy of a modern dual-socket server blade*

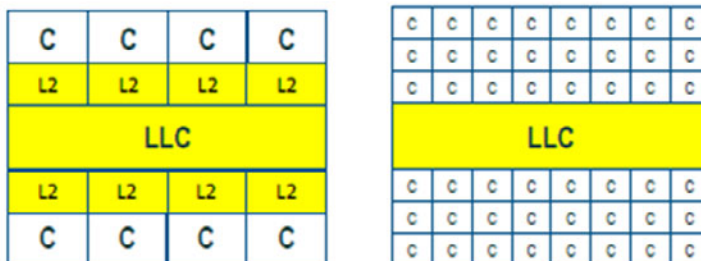


Figure 6.3. *Chip organization in an x86-based CPU (left) and a custom many-core CPU (right)*

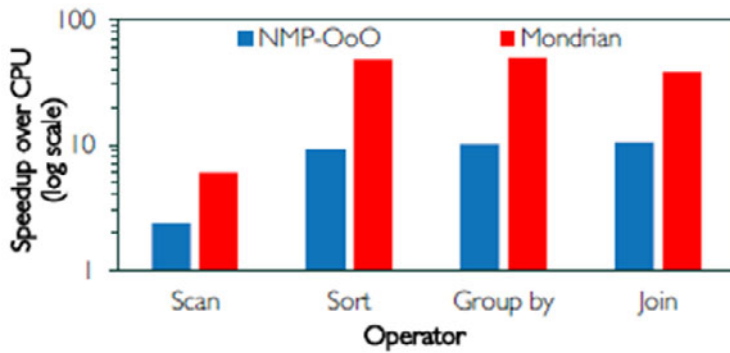


Figure 6.4. Speedup of near-memory processing: many-core OoO CPU and Mondrian with streaming algorithms

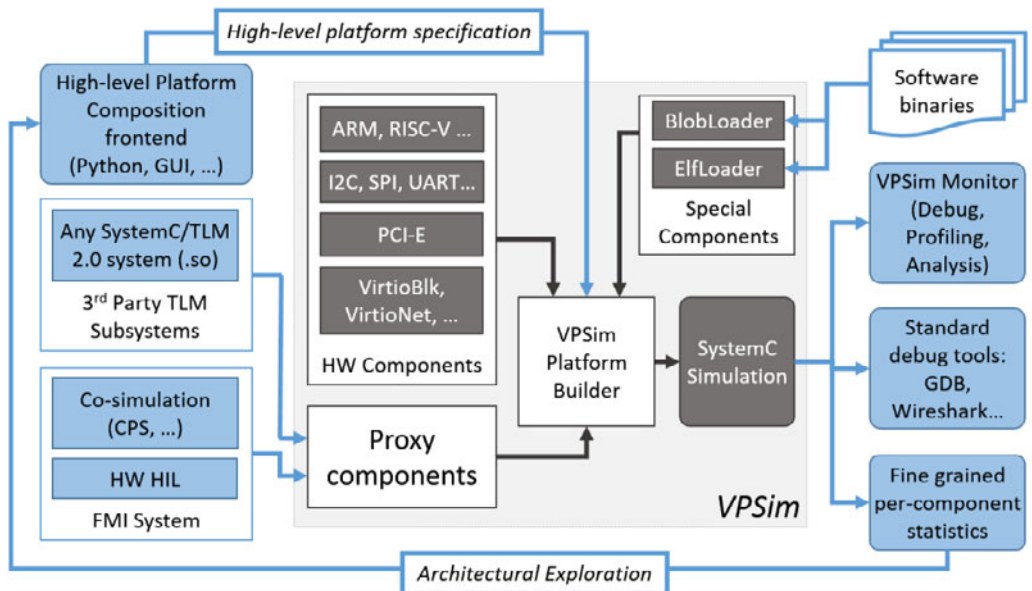


Figure 7.1. Overview on the operation of VPSim

```

1 import pyvp
2 import sys
3 import tty
4 import select
5 import os
6
7 class PL011Uart:
8     def __init__(self, **kwargs):
9         self.base=kwargs["base"]
10        self.irq=kwargs["irq"]
11        # ... other initializations
12
13        print("Initialized device pl011... (base=%s,irq=%s)" %
14              (self.base,self.irq))
15
16    def read(self, addr, size):
17        off = addr - self.base
18        self.update_status()
19        if off == 0x00:
20            return self.get_rx()
21        elif off == 0x18:
22            return self.get_status()
23        elif off == 0x3C:
24            return self.get_interrupt_unmasked()
25        elif off == 0x40:
26            return self.get_interrupt_masked()
27        else:
28            # other cases...
29
30    def write(self, addr, value, size):
31        off=addr-self.base
32        if off == 0x00:
33            self.backend_writechar(chr(value))
34        elif off == 0x38:
35            self.update_interrupt_enable(value)
36        elif off == 0x44:
37            self.upadte_interrupt_disable(value)
38        else:
39            # other cases...
40
41        self.update_status()
42        self.check_interrupt()
43
44    def check_interrupt(self):
45        if self.shouldInterrupt():
46            pyvp.interrupt(self, self.irq, 1)
47        else:
48            pyvp.interrupt(self, self.irq, 0)
49
50    def loop(self):
51        while True:
52            pyvp.wait(self, int(self.nsPerChar))
53            self.update_status()
54            self.check_interrupt()

```

Figure 7.2. *Simplistic implementation of the PL011 UART in Python*

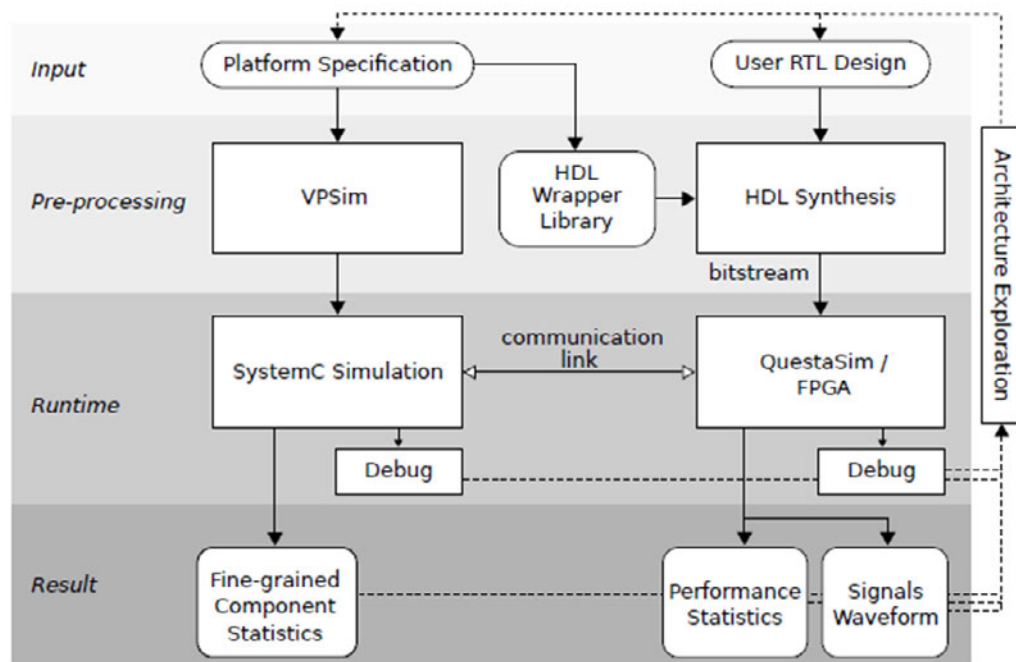


Figure 7.3. System validation flow using the hybrid prototyping solution

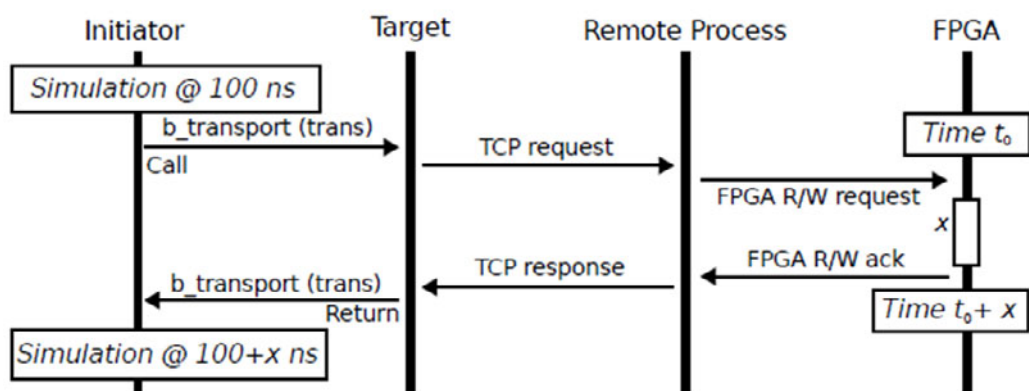


Figure 7.4. Synchronization mechanism between VPSim and FPGA in TLM R/W

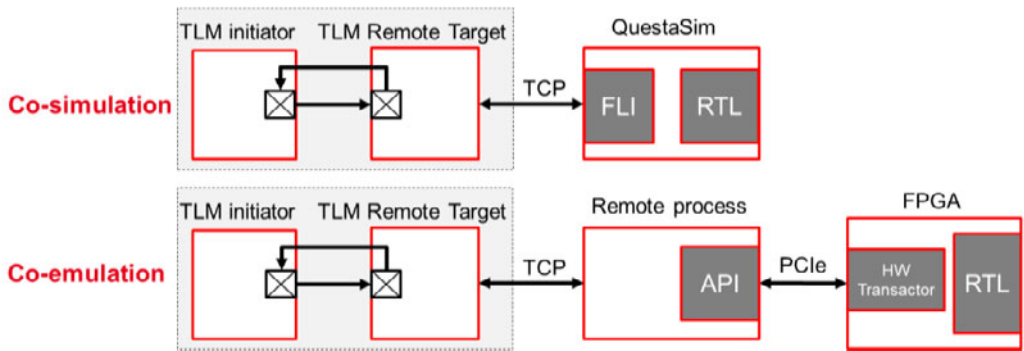


Figure 7.5. Communication scheme in VPSim during co-simulation and co-emulation

```

1  'can': [
2    {
3      'name': 'RenesasRcarH3CAN',
4      'base': 0xe6c30000,
5      'size': 0x1000,
6      'irq': 186,
7
8      'fmi_values': [
9        {
10         'name': 'speed',
11         'can_id': 0x10,
12         'causality': 'input',
13       },
14       {
15         'name': 'accel',
16         'can_id': 0x11,
17         'causality': 'output',
18       },
19       {
20         'name': 'brake',
21         'can_id': 0x12,
22         'causality': 'output',
23       },
24     ],
25   },
26 ]

```

Figure 7.6. Example of FmiValue declaration

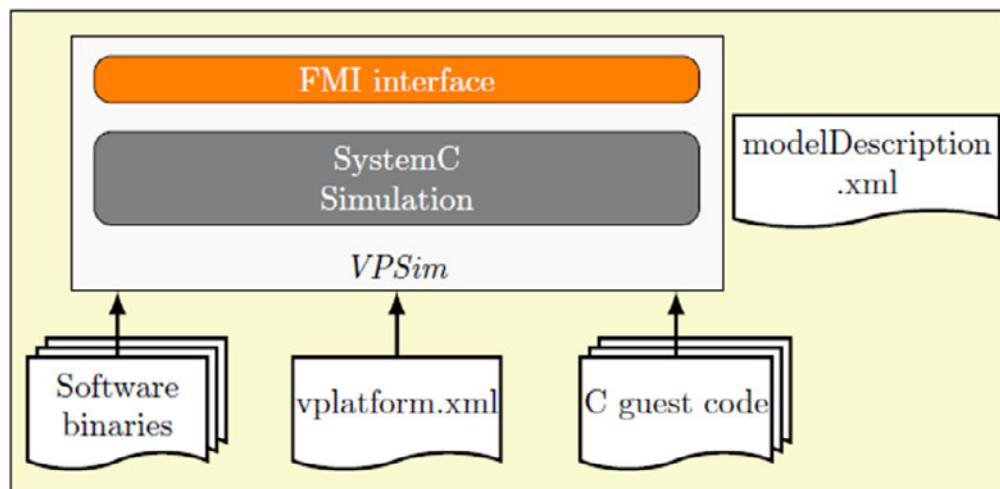


Figure 7.7. Structure of the generated virtual platform FMU

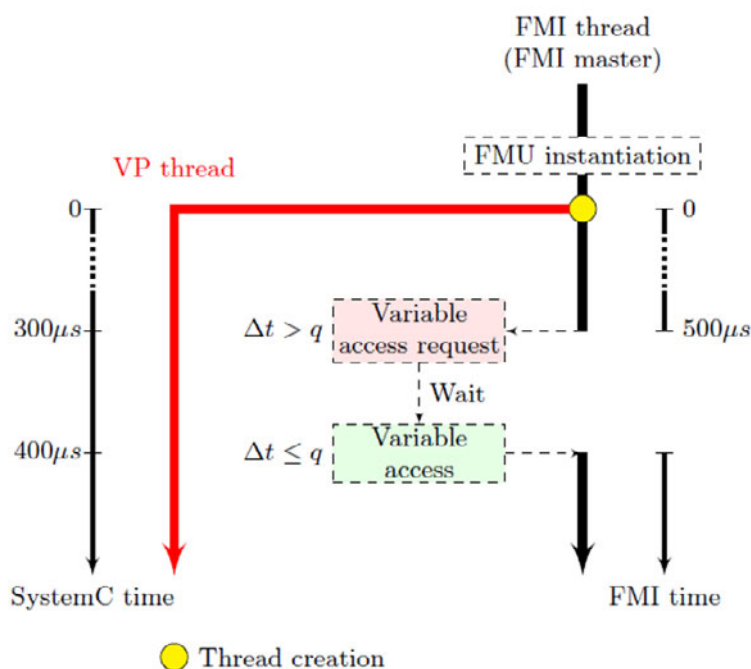


Figure 7.8. Parallel implementation of the virtual platform FMU

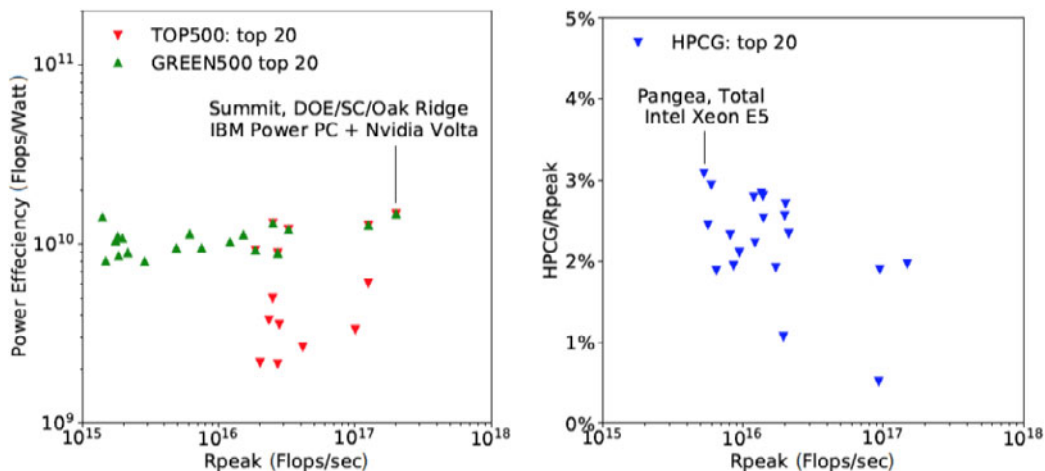


Figure 8.1. The top 20 of the TOP500 and GREEN500 supercomputers (left) and the top-20 of the HPCG supercomputers (right)

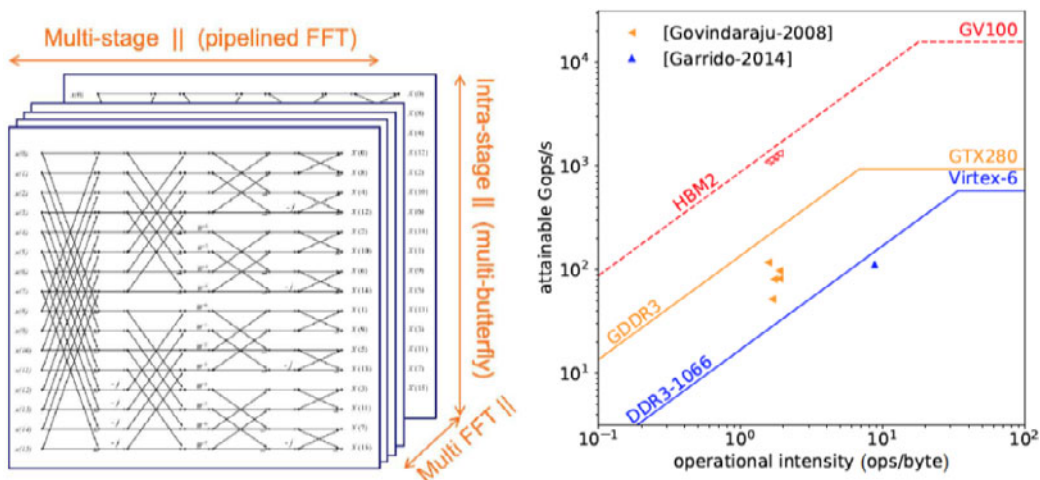


Figure 8.2. The three axes of FFT parallelism (left) and various machine rooflines with FFT performances

```

1 G = Graph()
2 G.n = [Node() for i in range(7)]
3 G.e = Path (G.n)
4 G.f = Edge (G.n[1], G.n[5])
5 G.e[1].init(D=3) # used only for the second simulation

```

Figure 8.3. Homogeneous flow graph G comprising seven nodes and seven edges

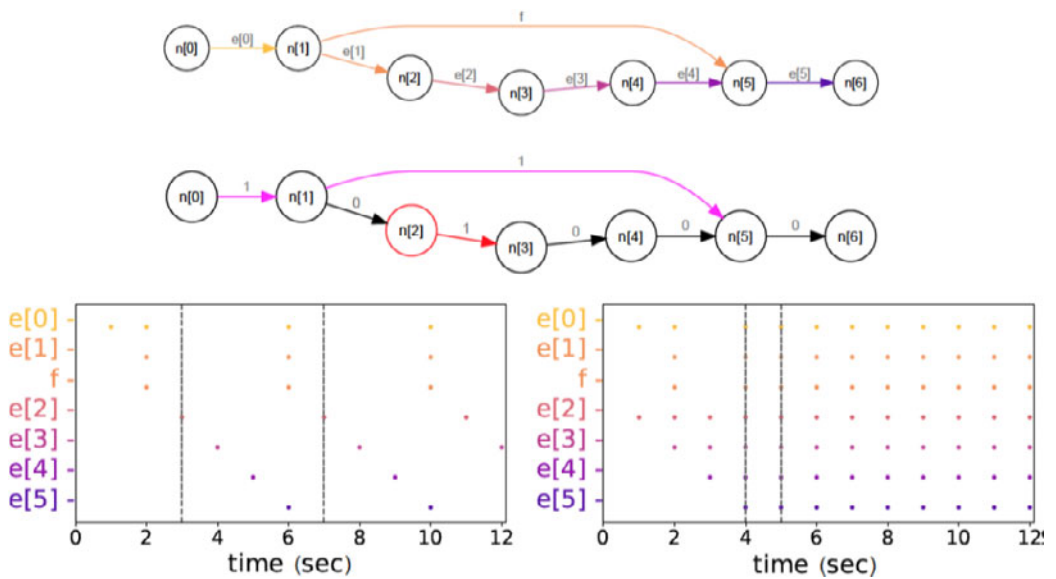


Figure 8.4. Flow graph G (top), its state after three cycles (mid), its flow (event timing) during 12 cycles (left) and its flow with three initial tokens on edge $e[1]$ (right)

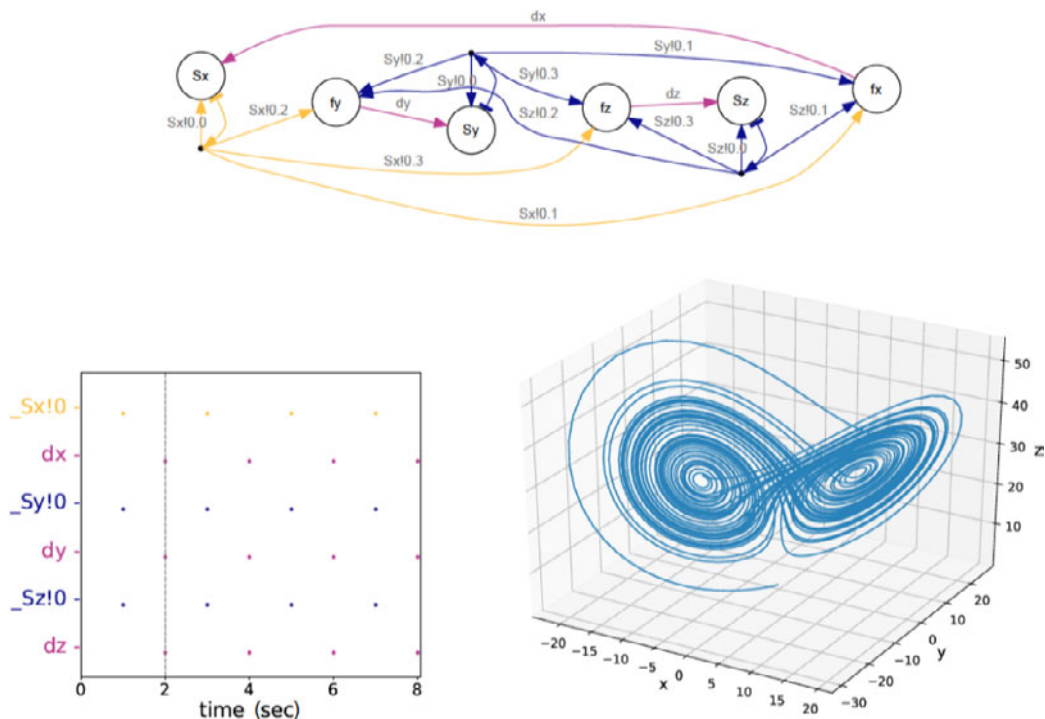


Figure 8.5. Graph Glorenz (top), its flow during the first eight cycles (left) and the x , y , z data produced by nodes f_x , f_y , f_z , respectively (right)

```

1 sigma,rho,beta, dt= 10, 28, 8/3. 0.01
2 fo      = [lambda x: sigma*(x[1] - x[0]),
3            lambda x: (rho-x[2])*x[0] - x[1],
4            lambda x: x[0]*x[1] - beta*x[2]]
5
6 Glorenz  = Graph(); G=Glorenz
7 G.Sx,G.Sy,G.Sz= [Node(I=0, fo= lambda x: x[0]+dt*x[1])
8                 for i in range(3)]
9 G.fx,G.fy,G.fz= [Node(I=[G.Sx.0[0],G.Sy.0[0],G.Sz.0[0]],fo=fo[i])
10                for i in range(3)]
11 G.dx,G.dy,G.dz= [Edge(G.fx,G.Sx),Edge(G.fy,G.Sy),Edge(G.fz,G.Sz)]
12
13 for i,n in enumerate([G.Sx, G.Sy, G.Sz]):
14     n.I[0].init(D=1, S=0, x=[-0.15, -0.2, 0.2][i])
15 for e in (G.dx, G.dy, G.dz): e.init(D=1, S=0, x=0)

```

Figure 8.6. Definition graph Glorenz


```

1 N      = 11
2 GMA    = Graph(rate=50e3); G= GMA
3 G.mp3  = MP3(mp3='../.../audio/war.mp3', tmin=37.5, rate=44100)
4 fo     = lambda x: np.int16((x[0][N]-x[0][0])/N +x[1])
5 G.ma   = Node(I=[G.mp3,0], fo=fo)
6 G.ma.I[0].init(x=[0 for i in range(N)])
7 G.ma.I[1].init(D=1, x=[0])           # accumulator, initialized to 0
8 G.snk  = Node(I=G.ma.O[0])

```

Figure 8.7. Definition of graph GMA

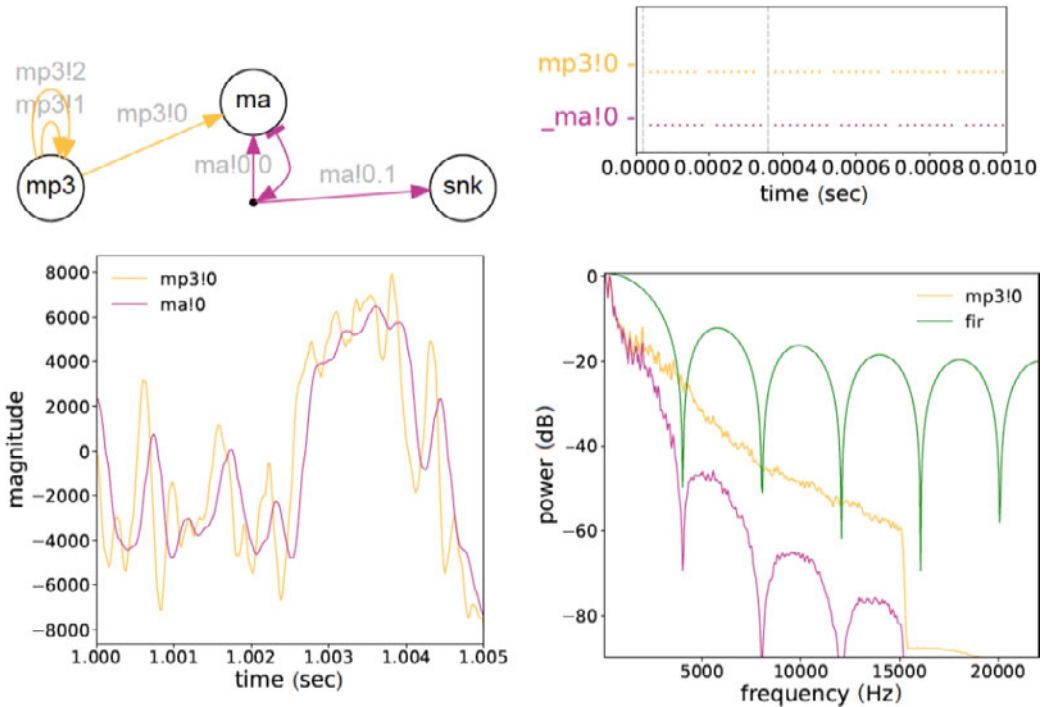


Figure 8.8. Node ma computes the moving average over a window $N = 11$ samples. The flow shows incidental gaps due to flow throttling by the mp3 source. The two bottom plots show the signal smoothing in time and frequency domain

```

1 set_fsm([Repeat(L, iter=Rule(I=[1,0,1], O=[0,1,1], s=0), exit=1),
2         Fire(Rule(I=[1,1,1], O=[1,1,1], s=1))])

```

Figure 8.9. An example of an FSM for a cyclo-static dataflow node

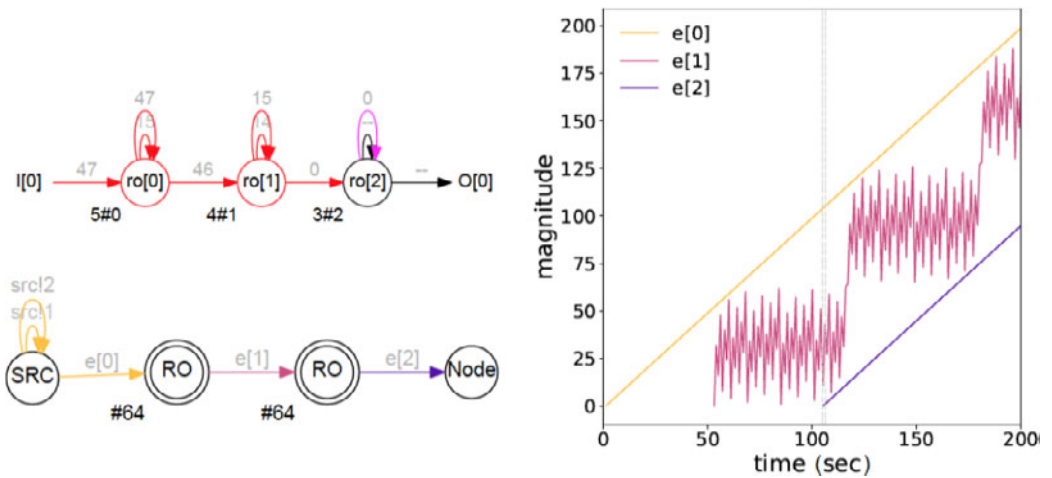


Figure 8.10. Subgraph RO (left, top) comprises three CSDF nodes $ro[i]$ and performs an $N = 64$ FFT bit-reversal. As FFT bit-reversal is its own inverse, a cascade of two ROs performs the identity, as shown for a linear input sequence (right)

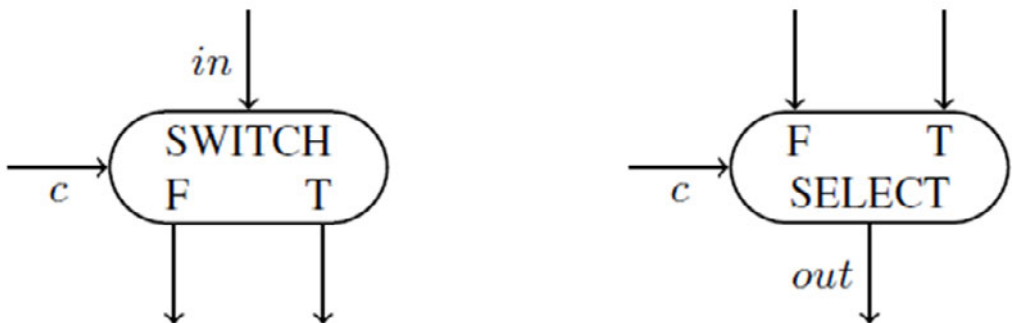


Figure 8.11. SWITCH and SELECT nodes as used in Boolean Dataflow

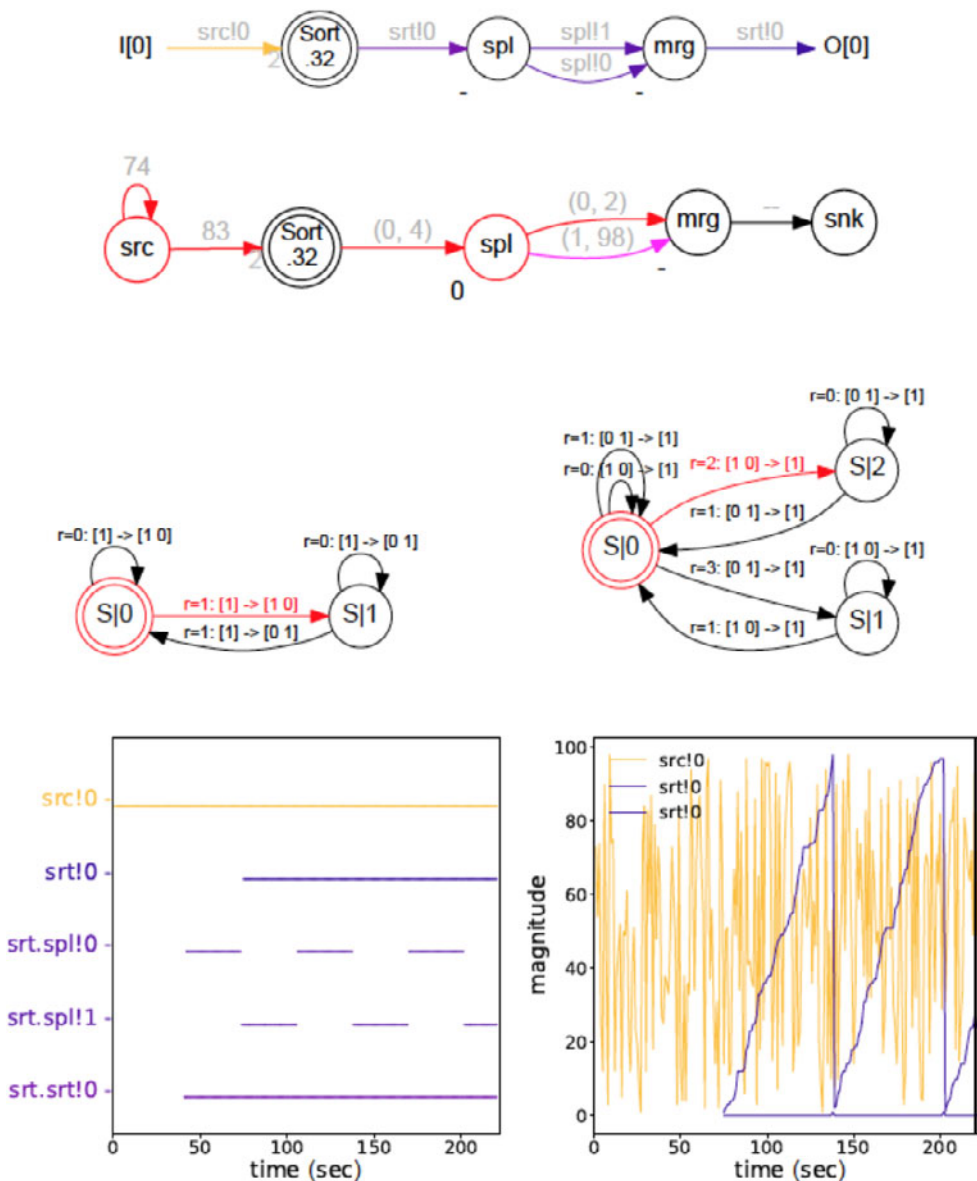


Figure 8.12. Recursive dataflow graph Sort.64 (top), Sort.64 connected to a random-number source `src` and `snk` (one-from-top), the FSMs of nodes `spl` and `mrq` (mid-left and right) and the flow and data plots for Sort.64 (bottom)

```

1 class Split(Node):
2     def __init__(self, N, I=[]):
3         super(Split, self).__init__(I=I)
4         pass0 = lambda s: Rule(I=(1,), O=(1,0), s=s)
5         pass1 = lambda s: Rule(I=(1,), O=(0,1), s=s)
6         self.set_fsm([Select([pass0(0), pass0(1)]),
7                       Select([pass1(1), pass1(0)])])
8         self.set_fs([lambda x: x[0], lambda x: x[0]])
9         self.set_fo([lambda x: x, lambda x: x])

```

Figure 8.13. Definition of node class Split

```

1 class IMG(Node):
2     def __init__(self, N I=[]):
3         super(IMG, self).__init__(I=I)
4         rules= [Rule(I=[int(j==i) for j in range(N)],O=[1,],s=0)
5                 for i in range(N)]
6         self.set_fsm([Choice(rules)])
7         self.set_fo([lambda x, r: x[r]])

```

Figure 8.14. Definition of node class IMG

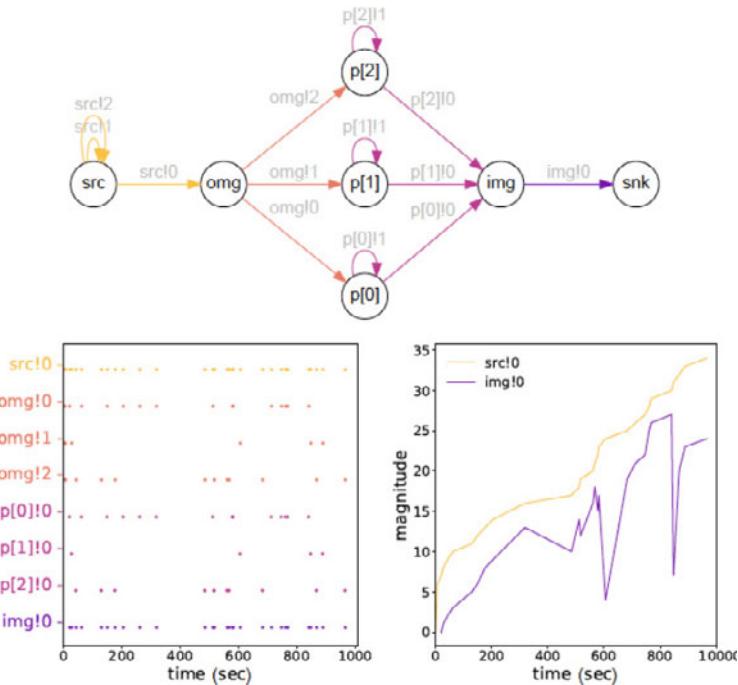


Figure 8.15. A toy processor farm: dataflow graph (top), flow plot (left) and data plot (right). Node `src` produces the natural numbers, in order

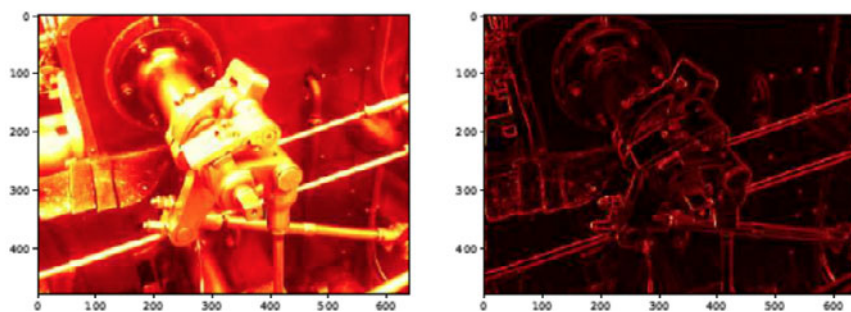
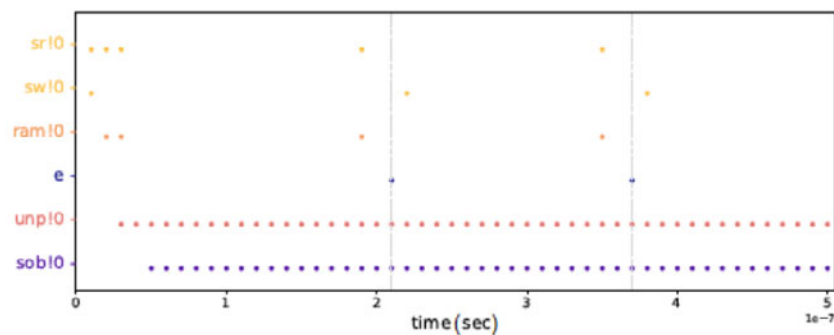
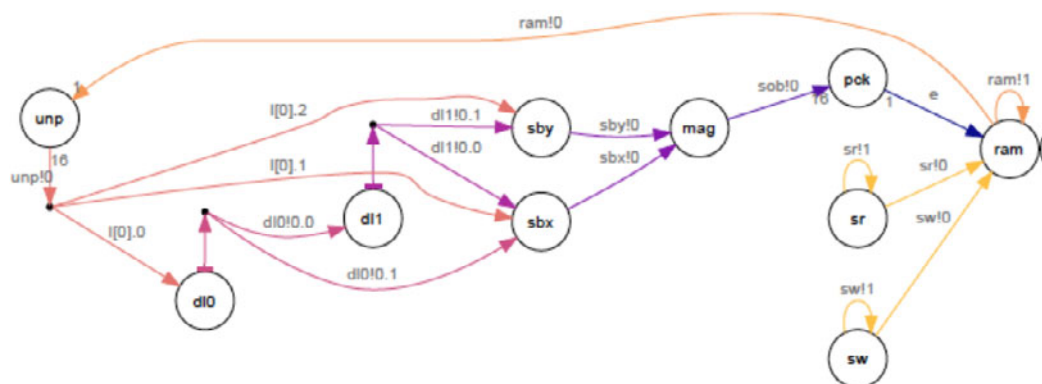


Figure 8.16. Dataflow graph Sobel with I/O to RAM node ram (top), 50 μ sec flow plot (middle) and example input+output images (bottom)

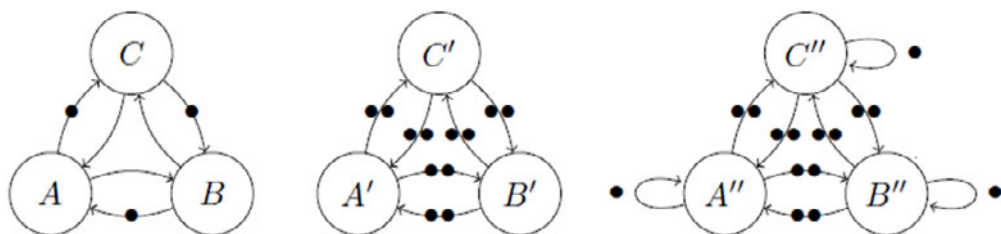


Figure 8.17. Three tricycle dataflow graphs. Node execution times equal 1

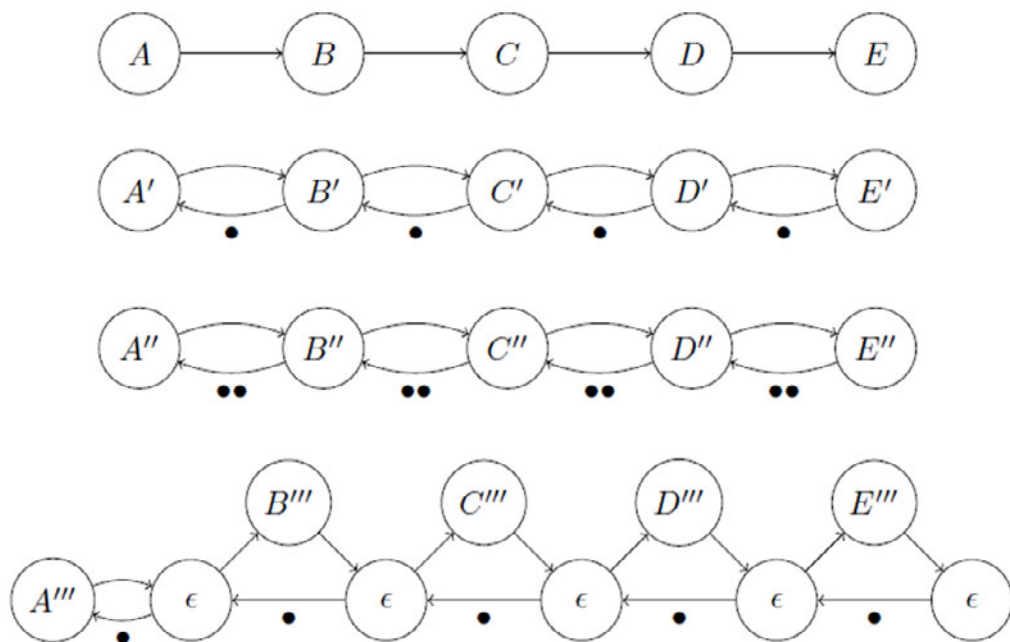


Figure 8.18. A dataflow pipeline with four different back-pressure schemes. Nodes A to E have an execution time of 1 and self-loops to bound the node's throughput to 1 (not drawn). ϵ -nodes have an execution time of $\epsilon \rightarrow 0$

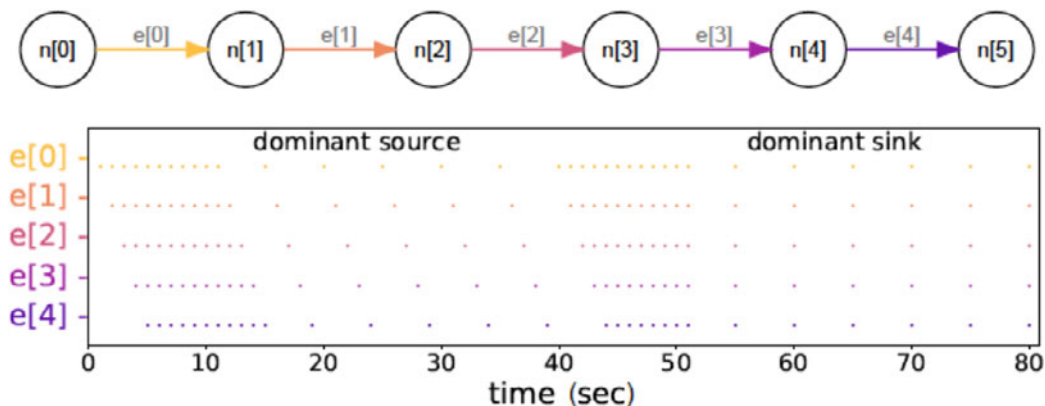


Figure 8.19. A six-node pipeline starts with all nodes running self-timed with $t_n = 1$. After 11 firings, source node $n[0]$ is throttled to $t_{n[0]} = 5$. At time=40, the source throttle is switched off, and at time=50, the sink node is throttled to $t_{n[5]} = 5$

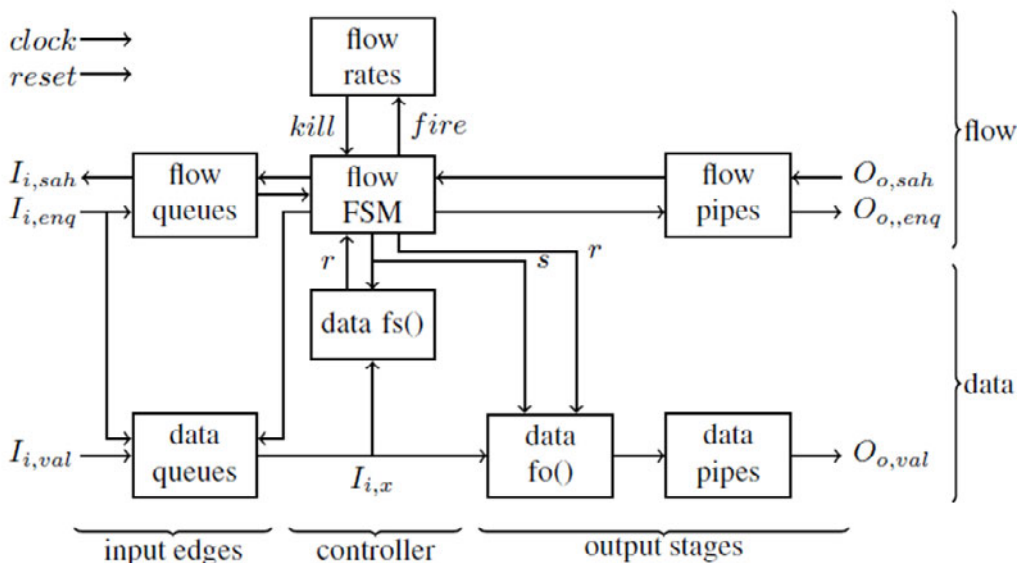
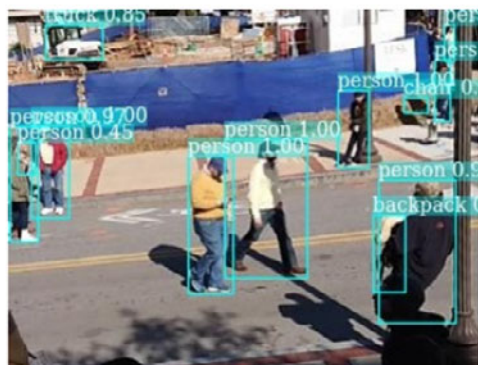


Figure 8.20. Verilog template for a StaccatoLab node, including input edges



visible



infrared

Figure 9.1. Visible and infrared images and associated classification from the CAMEL dataset (<https://camel.ece.gatech.edu>)

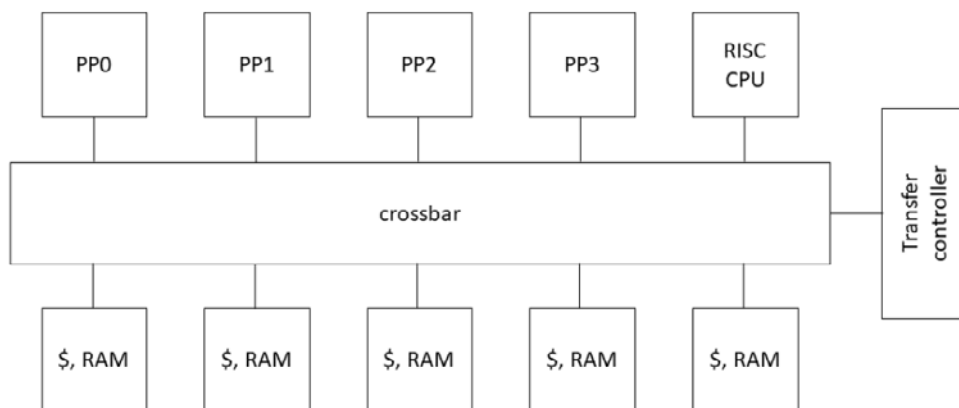


Figure 9.2. The TI MVP

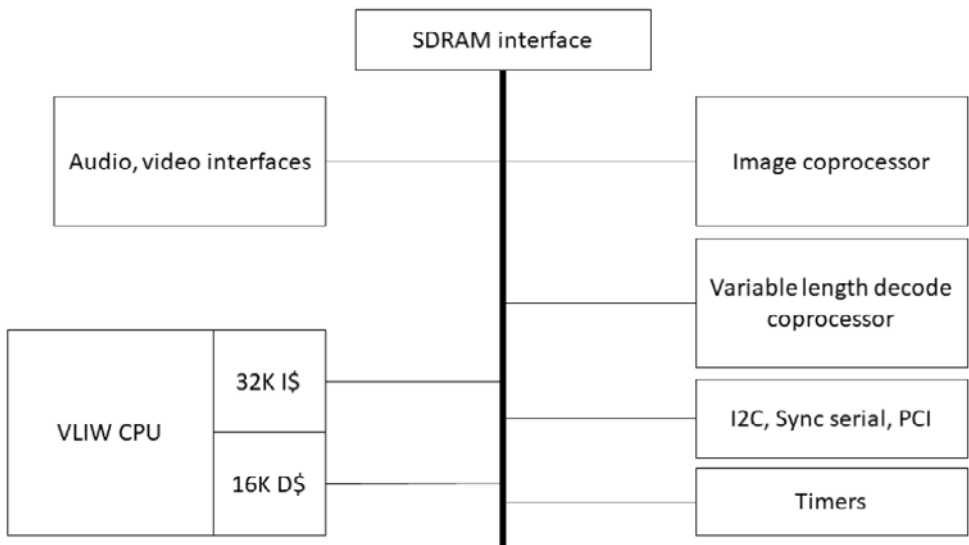


Figure 9.3. *The Trimedia TM-1*

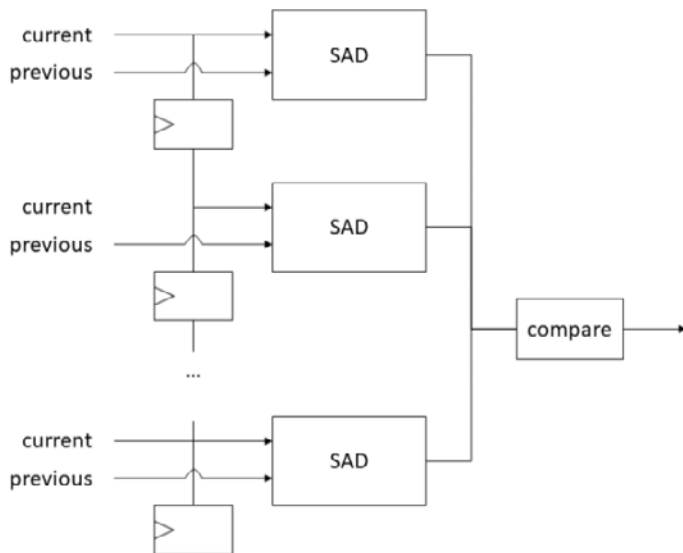


Figure 9.4. *Motion compensation accelerator*

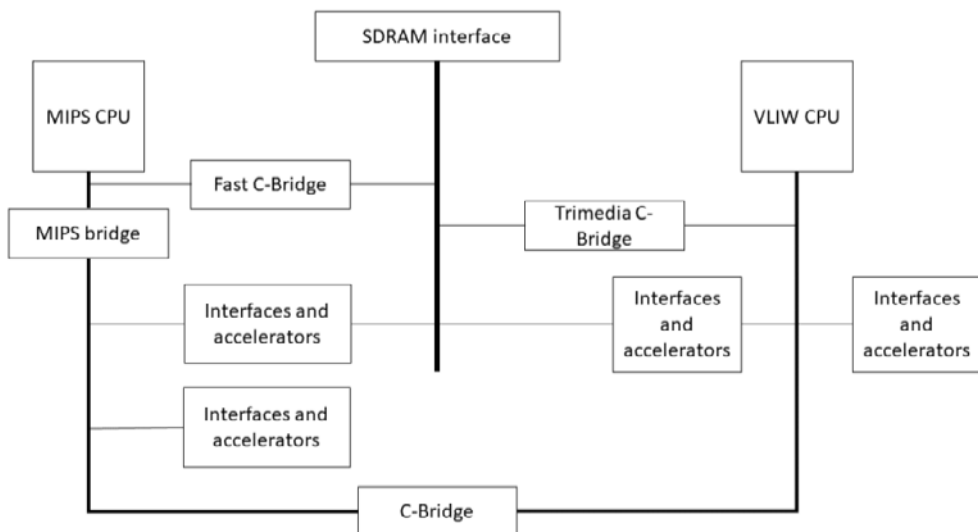


Figure 9.5. *The Philips Viper*

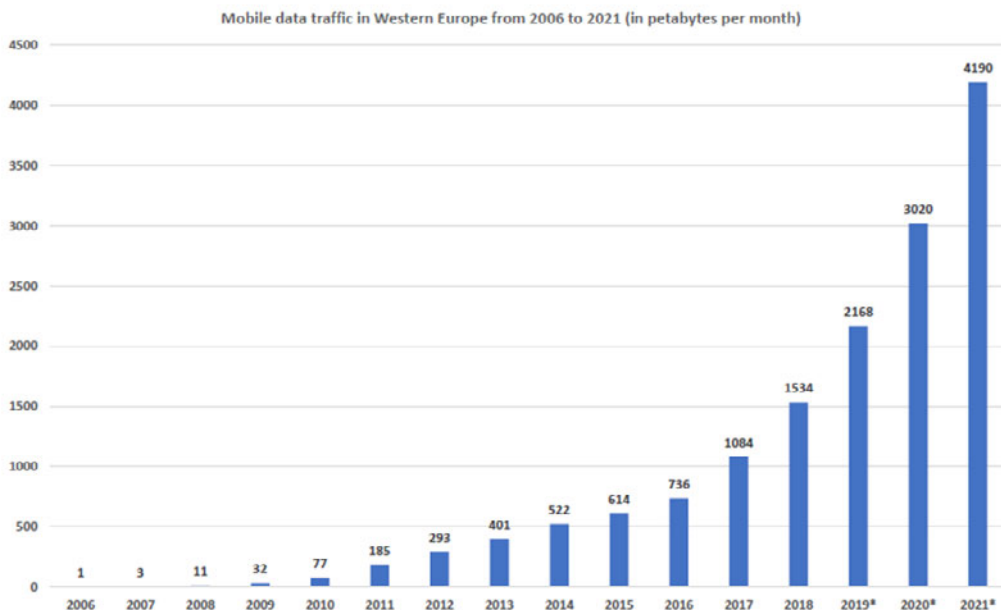


Figure 10.1. *Mobile data traffic in western Europe from 2006 to 2021 (2019-2021 are projections), adapted from Page et al. (2010) and Cisco Systems (2017)*

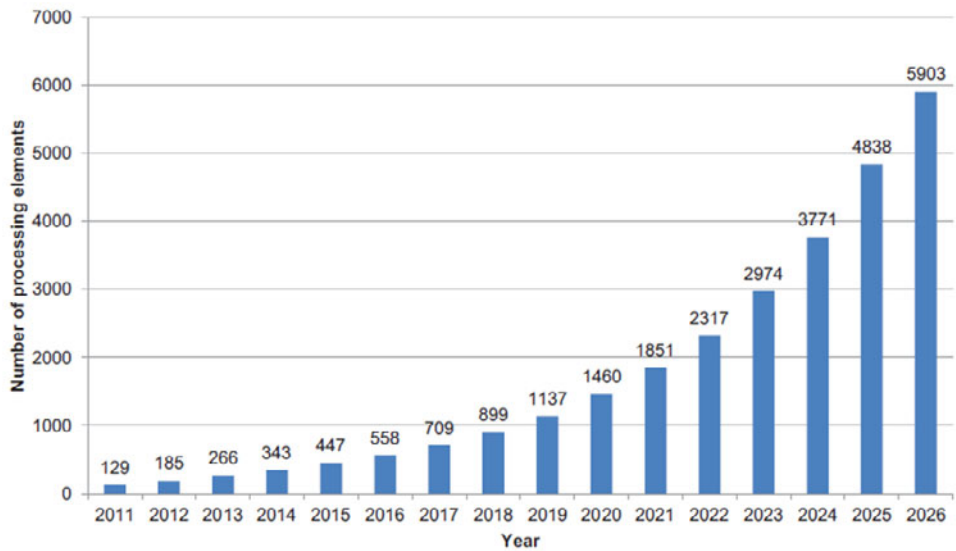


Figure 10.2. SOC consumer portable design trends (adapted from International Technology Roadmap for Semiconductors (2011)). The processing elements are referred to cores or parallel execution units in CPUs, GPUs, accelerators, etc.

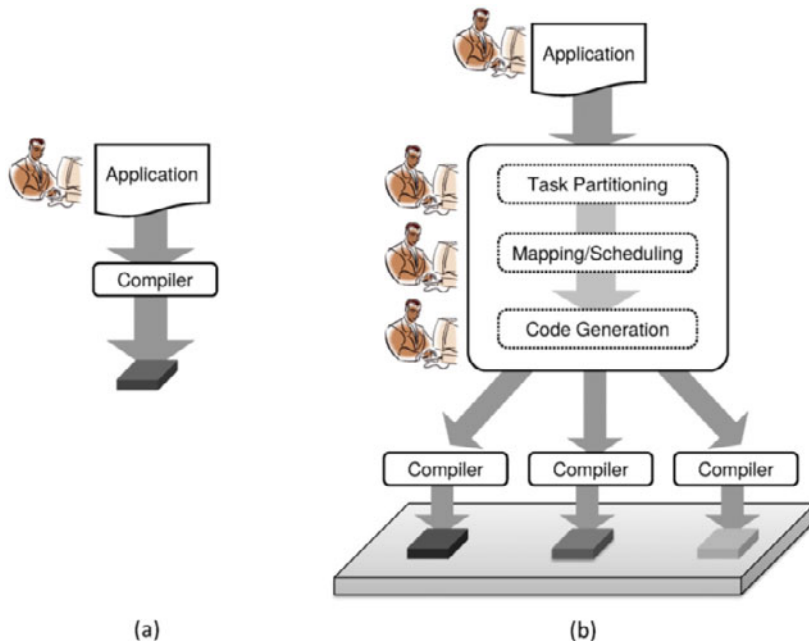


Figure 10.3. Programming flow. (a) Uni-processor: compilers perform an end-to-end translation from parsing source code, target independent optimization and targetdependent optimization to final binary generation. (b) MPSoC: no compilation framework for MPSoCs is in place. The process (task partitioning, mapping/scheduling and code generation) is manual

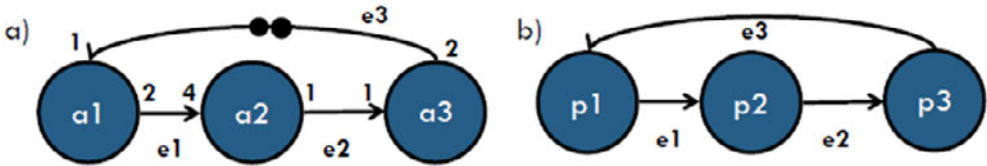


Figure 10.4. a) Example of an SDF graph. Actors have fixed consumption and production rates. Initial tokens are represented as dots on the channels. b) Example of a process network. The abstract specification expresses that nodes can communicate if there is a channel that connects them, but does not specify how this communication takes place

```

1 typedef struct { int i; double d; } my_struct_t;
2 typedef union { float f; short s[4]; } my_union_t;
3
4 __PNchannel char B[3][3];
5 __PNchannel my_struct_t C;
6 __PNchannel my_union_t D;
7 __PNchannel int A = {1, 2, 3}; /* Initial channel tokens */

```

Figure 10.5. CPN example code: channel declaration

```

1 /* Run Length Decoding, e.g. 4A2B5C3D -> AAAABBBCCCCDDDD */
2 __PNkpn RLD __PNin(int EncIn) __PNout(int DecOut)
3 {
4     int count, i;
5     while (1) {
6         __PNin(EncIn) /* read a token (# of appearances) from EncIn,
7                        e.g. 4 */
8         count = EncIn;
9         __PNin(EncIn) /* read a token (data itself) from EncIn,
10                       e.g. A */
11         for (i = 0; i < count; ++i) /* write data to DecOut,
12                                     e.g. 4 times of A */
13             __PNout(DecOut)
14             DecOut = EncIn;
15     }
16 }

```

Figure 10.6. CPN example code: KPN process template (run-length decoding)

```

1 __PNsdf Add __PNin(int a, int b) __PNout(int sum) {
2     /* initialization code could be placed here */
3     __PNloop { /* infinite loop */
4         /* a and b are read from the channel implicitly */
5         sum = a + b; /* channel variables are accessible in C code */
6         /* sum is written to the channel implicitly */
7     }
8 }

```

Figure 10.7. CPN example code: SDF process template (adder)

```

1 __PNchannel int decoder1_input = {4, 'A', 2, 'B', 5, 'C', 3, 'D'};
2 __PNchannel int decoder2_input = {3, 'E', 5, 'F', 4, 'G', 2, 'H'};
3
4 __PNchannel int decoder1_output, decoder2_output;
5 __PNchannel int add_output;
6
7 __PNprocess decoder1 = RLD __PNin(decoder1_input)
8     __PNout(decoder1_output);
9 __PNprocess decoder2 = RLD __PNin(decoder2_input)
10    __PNout(decoder2_output);
11 __PNprocess add = Add __PNin(decoder1_output, decoder2_output)
12    __PNout(add_output);

```

Figure 10.8. CPN example code: process instantiation

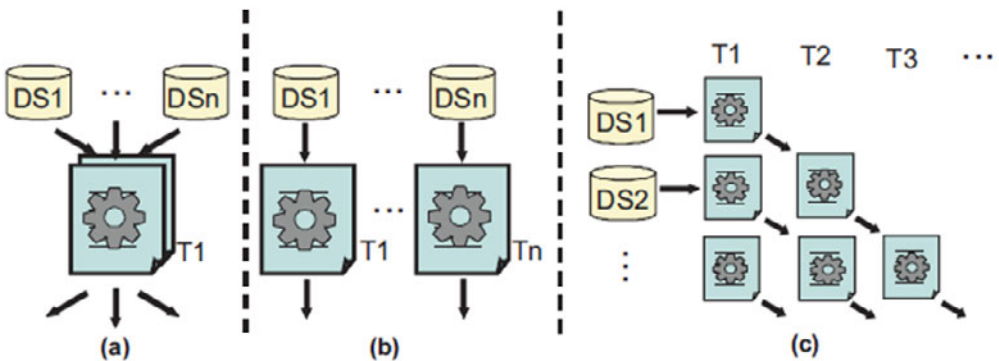


Figure 10.9. Types of parallelism: (a) DLP, (b) TLP and (c) PLP
(DS stands for data set and T stands for task)

```

1 __PNchannel int DS1, DS2, DSN;
2 __PNchannel int Result1, Result2, ResultN;
3 __PNprocess T_1 = SameTaskT1 __PNin(DS1) __PNout(Result1);
4 __PNprocess T_2 = SameTaskT1 __PNin(DS2) __PNout(Result2);
5 __PNprocess T_N = SameTaskT1 __PNin(DSN) __PNout(ResultN);

```

Figure 10.10. CPN example code: DLP

```

1 __PNchannel int DS1, DS2, DSN;
2 __PNchannel int Result1, Result2, ResultN;
3 __PNprocess T_1 = Task1 __PNin(DS1) __PNout(Result1);
4 __PNprocess T_2 = Task2 __PNin(DS2) __PNout(Result2);
5 __PNprocess T_N = TaskN __PNin(DSN) __PNout(ResultN);

```

Figure 10.11. CPN example code: TLP

```

1 __PNchannel pixels DS, intermediate1, intermediate2,
  intermediate3, output;
2 __PNprocess source = ReadImage __PNout(DS);
3 __PNprocess T_1 = PipelineTask1 __PNin(DS)
  __PNout(intermediate1);
4 __PNprocess T_2 = PipelineTask2 __PNin(intermediate1)
  __PNout(intermediate2);
5 __PNprocess T_3 = PipelineTask3 __PNin(intermediate2)
  __PNout(intermediate3);
6 __PNprocess T_N = PipelineTaskN __PNin(intermediate3)
  __PNout(output);
7 __PNprocess sink = WriteImage __PNin(output);

```

Figure 10.12. CPN example code: PLP

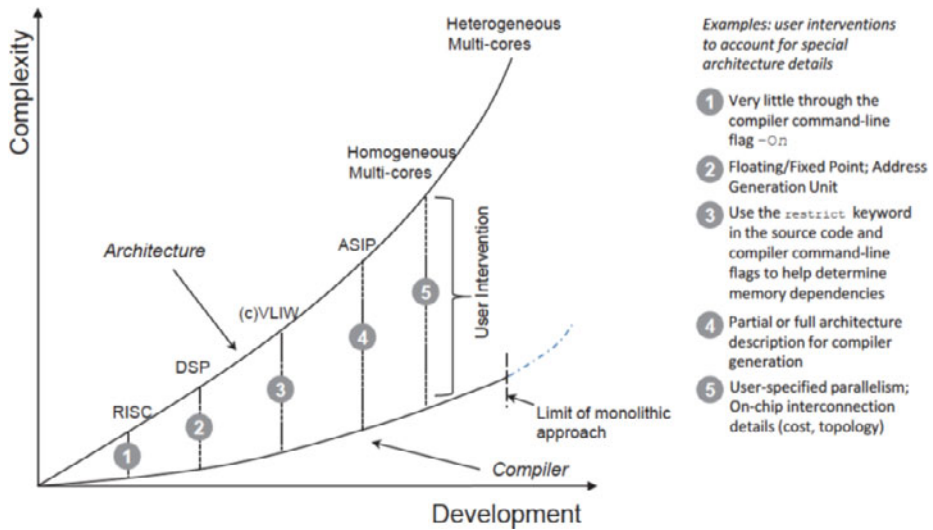


Figure 10.13. Evolution of compiler development versus architecture development

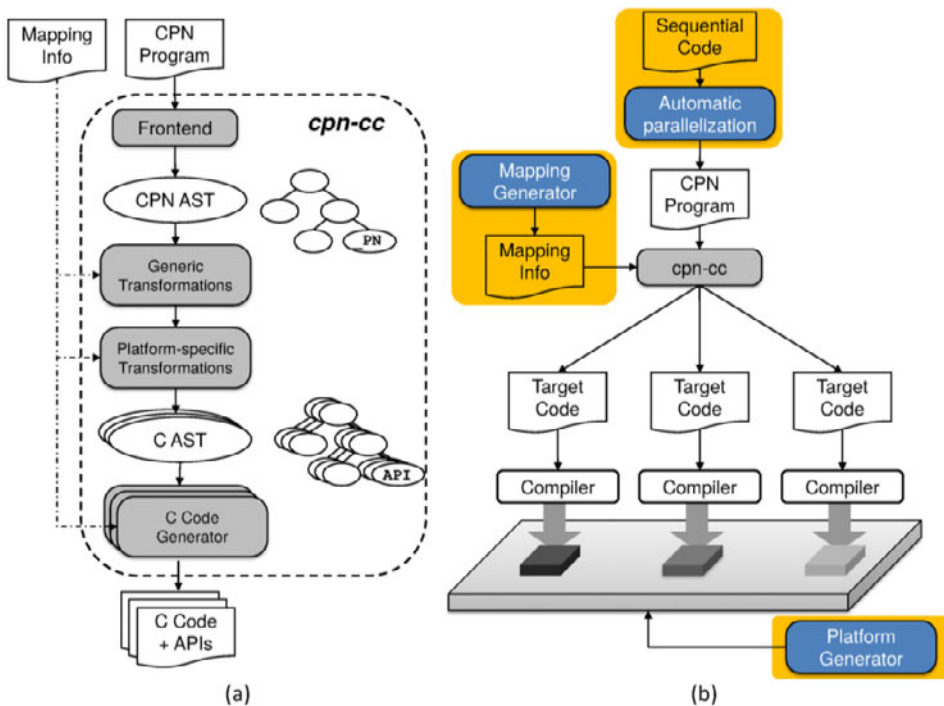


Figure 10.14. (a) Structure of the source-to-source compiler *cpn-cc*. (b) An example of a complete compiler framework for a three-processor heterogeneous MPSoC: the parts with a shaded background are optional to the core compiler framework

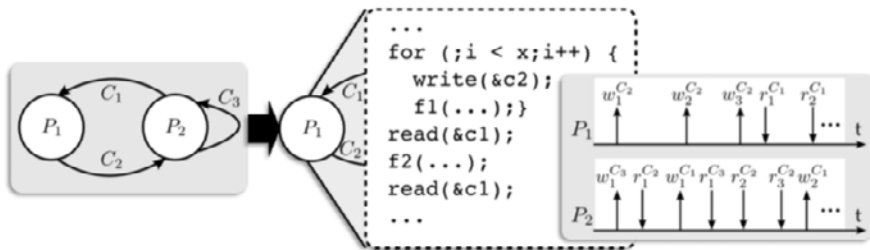


Figure 10.15. Illustration of process traces. Example process network with a sample code of the logic of process P_1 , alongside a hypothetical trace of read and write events that may result from the process inner control flow

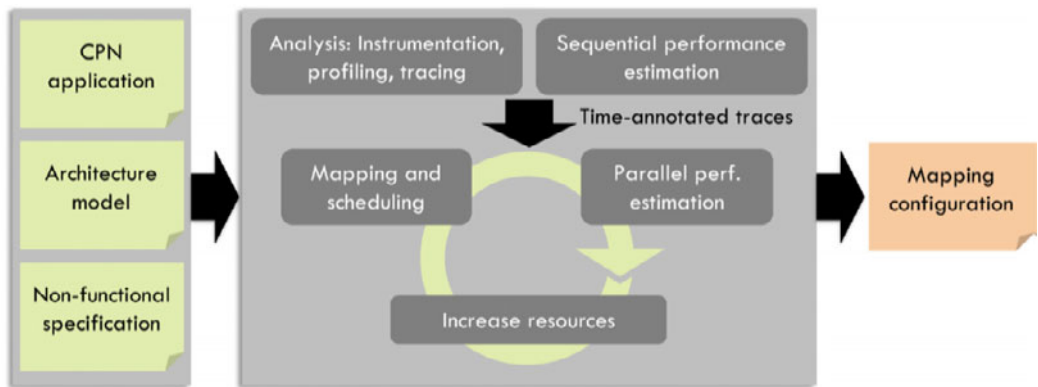


Figure 10.16. Overview of the mapping flow

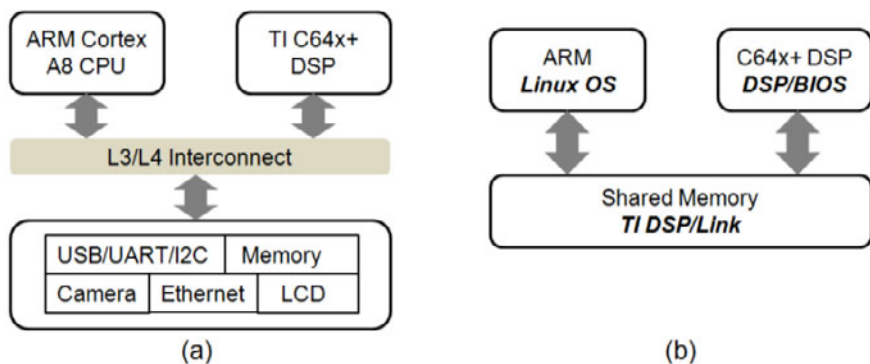


Figure 10.17. (a) OMAP3530 block diagram. (b) Overview of TI OMAP software

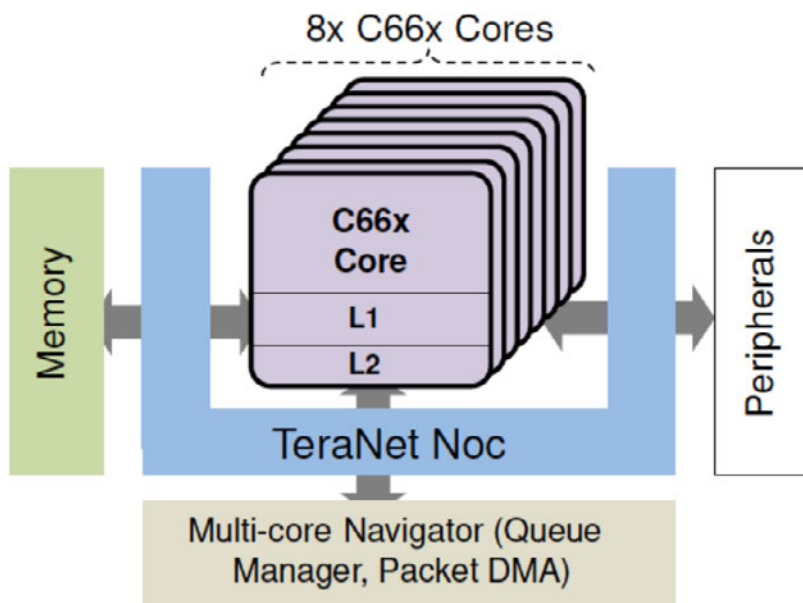


Figure 10.18. TI C6678 block diagram

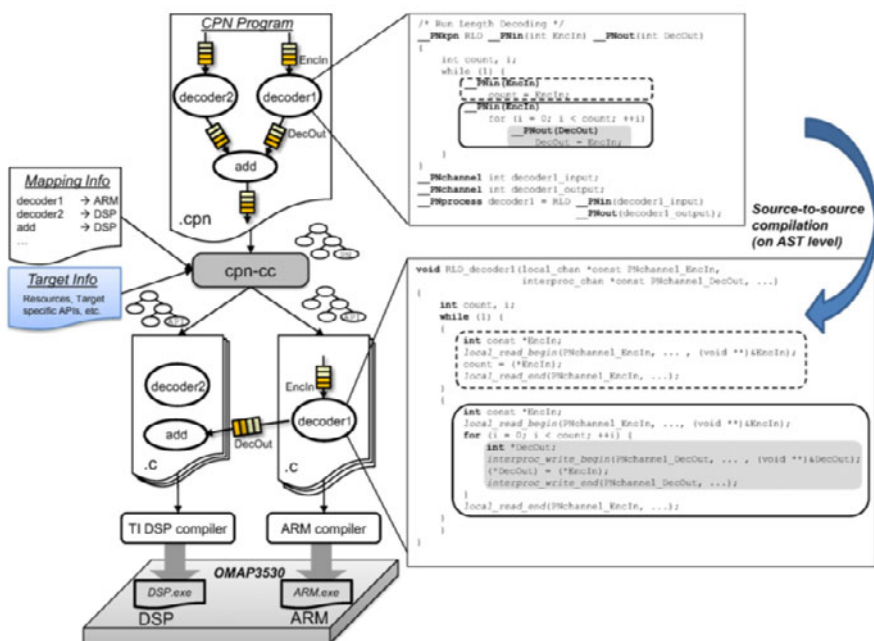


Figure 10.19. A complete compilation framework for OMAP3530 using MAPS

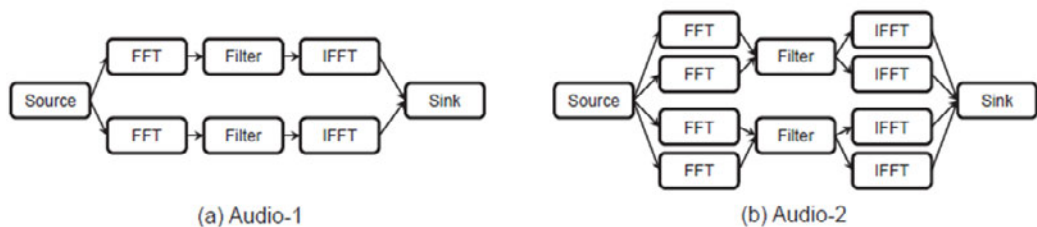


Figure 10.20. Benchmark: digital audio filter

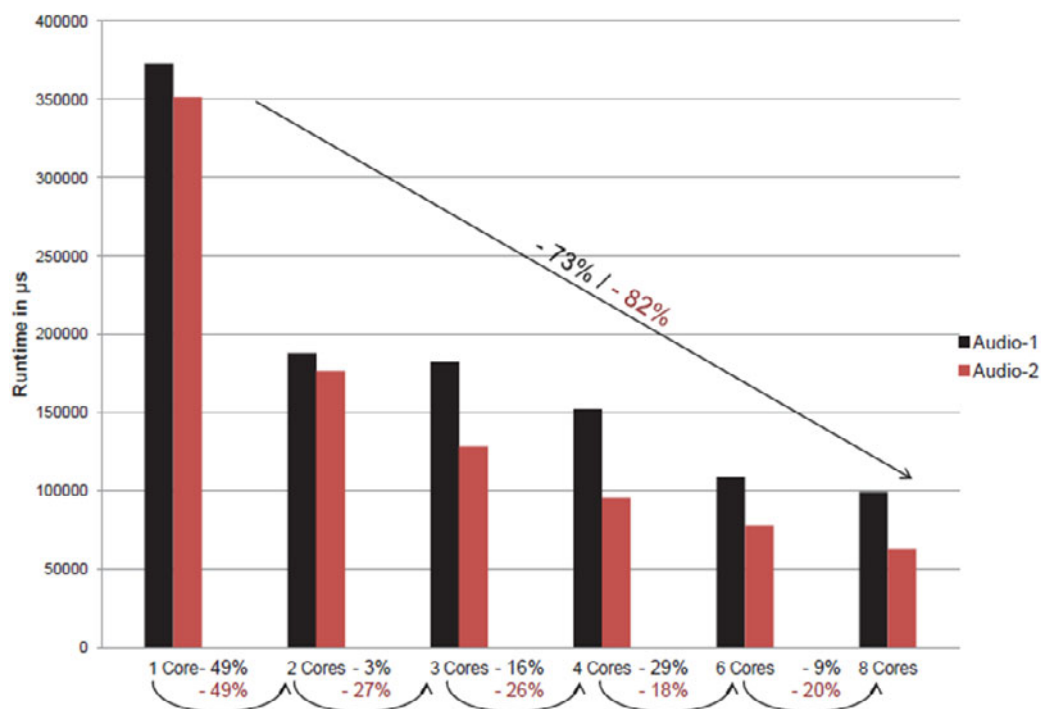


Figure 10.21. Results of the audio filter (the percentage indicates the reduction of runtime between two mappings)

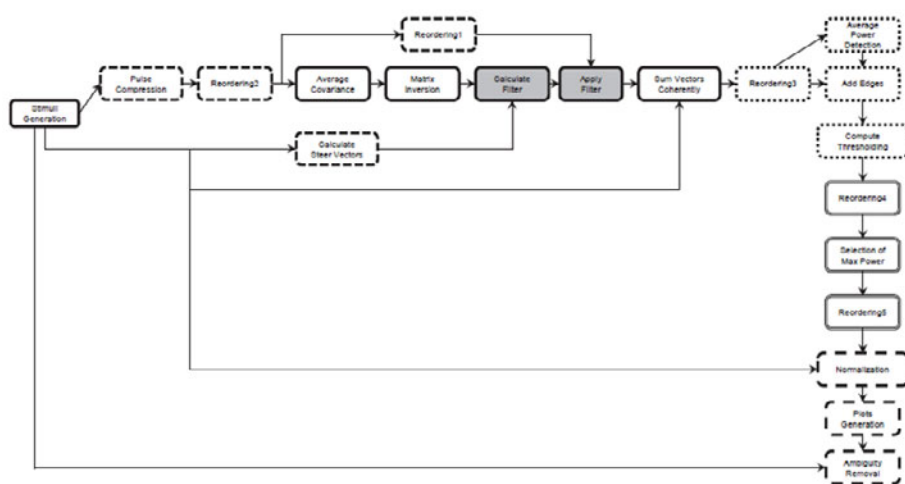


Figure 10.22. Benchmark: radar application

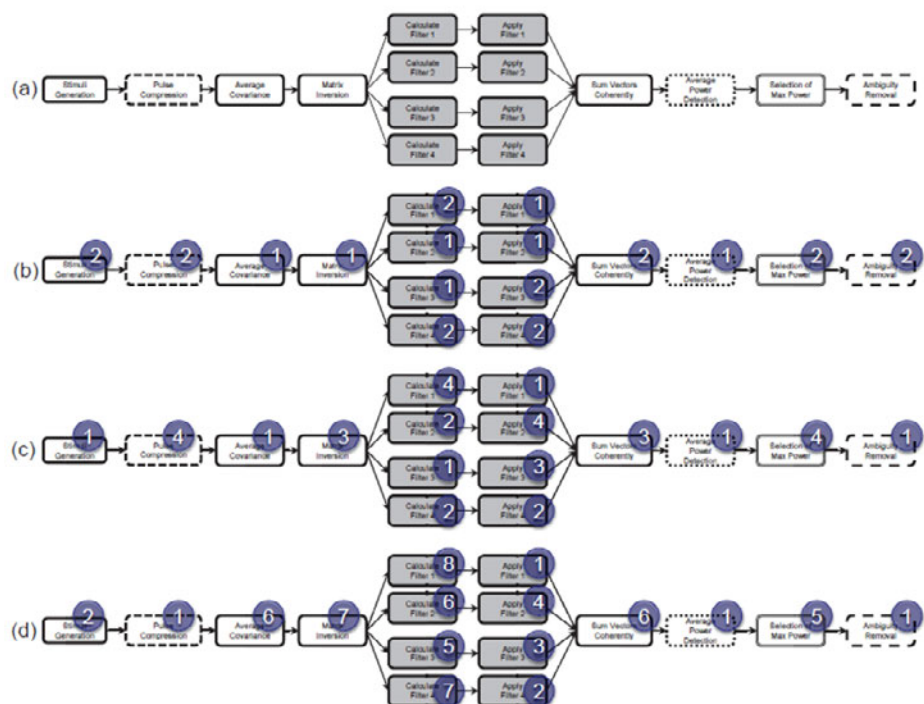


Figure 10.23. Mapping results of the radar application: improved version. (a) Process network topology (only edges that communicate major data traffic are plotted for the sake of clarity). (b) Mapping onto two cores. (c) Mapping onto four cores. (d) Mapping onto eight cores

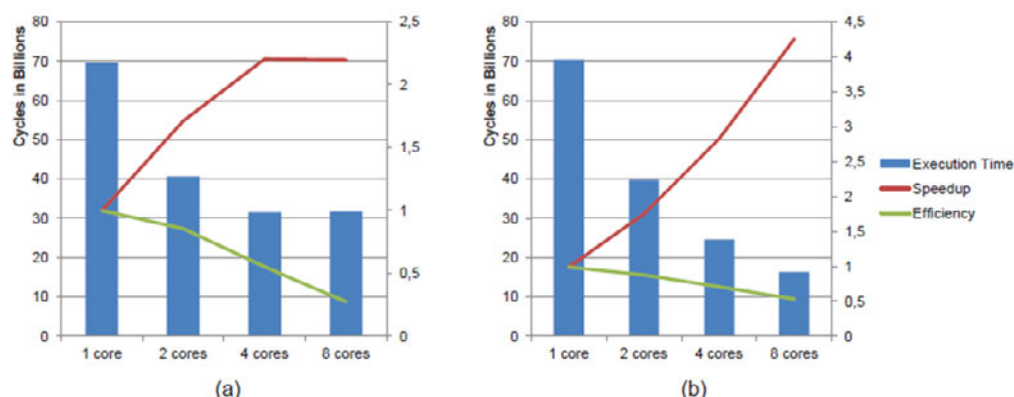


Figure 10.24. Results of the radar applications: performance. (a) Initial version and (b) improved version. $Speedup = \frac{Execution\ Time}{Sequential\ Time}$; $Efficiency = \frac{Speedup}{Number\ of\ cores\ used}$