

Computer Generation of Efficient Software Viterbi Decoders

**Frédéric de Mesmay, Srinivas Chellappa,
Franz Franchetti, Markus Püschel**

**Electrical and Computer Engineering
Carnegie Mellon University**

**Co-Founder
SpiralGen, Inc.**

Sponsors: DARPA DESA program, ONR, NSF-NGS/ITR, NSF-ACR, Mercury, and Intel

Viterbi Decoder

■ Error correction

- Forward Error Correction
- Digital cellular (CDMA, GSM), modems, satellite/deep space communications, 802.11 wireless LANs
- Software defined radio (SDR)



GSM (TCH/FS)

K=5 rate=1/2

CDMA2000/UMTS/IS-95

K=9 rate=1/3

■ Pattern Recognition

- Speech recognition
- text recognition
- computational linguistics
- bioinformatics



NASA Cassini

Orbiter:

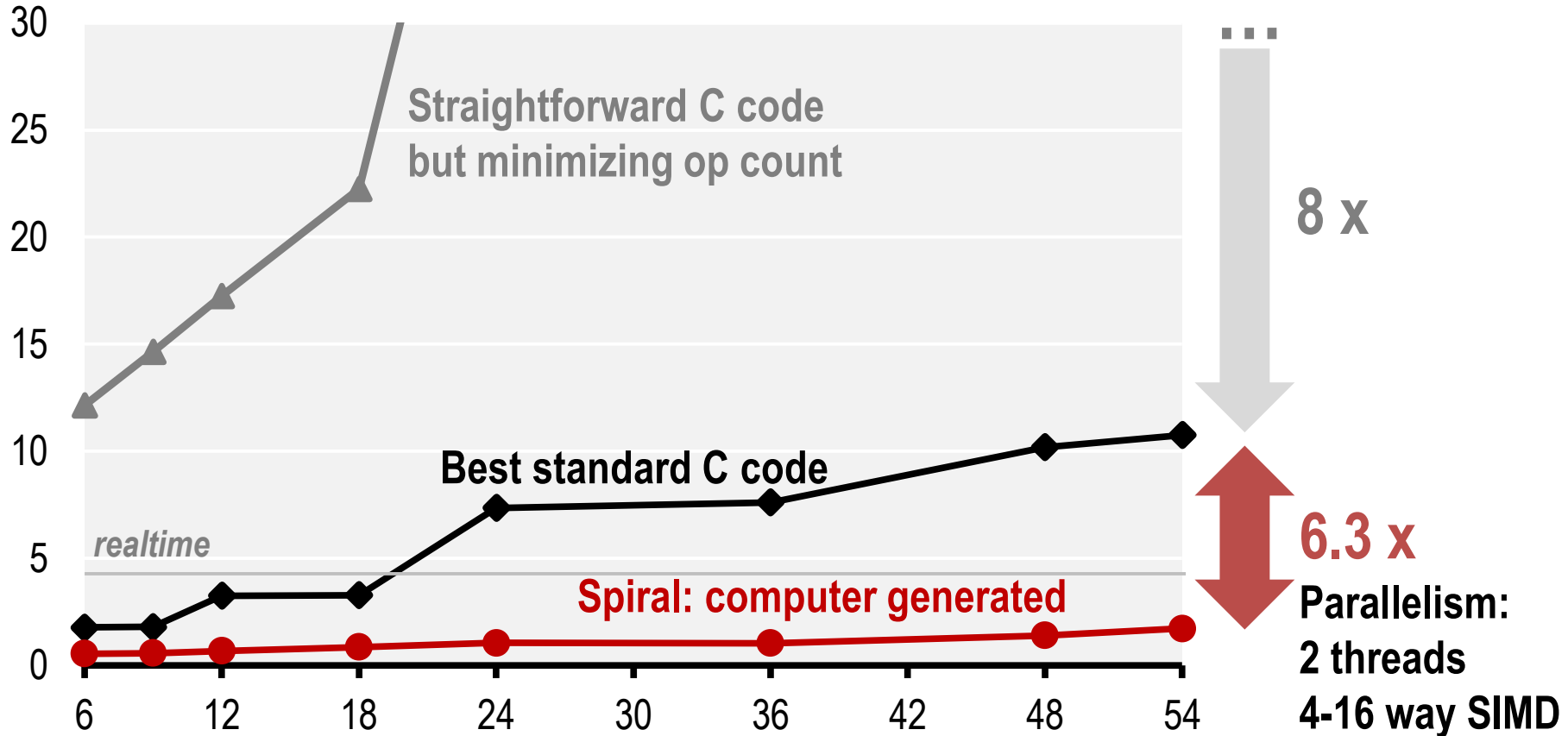
K=15 rate=1/6

SDR requires efficient Viterbi decoder software implementations

Software Defined Radio

WiFi transmitter on Intel Atom Dualcore

Run time per OFDM symbol [μ s] vs. data rate [Mbit/s]



Compilers fail to optimize: 50x

Spiral: Viterbi Software Generation

Select convolutional code

Select a preset code or customize parameters

custom

Voyager

NASA-DSN

CCSDS/NASA-GSFC

WiMax

CDMA IS-95A

LTE (3GPP - Long Term Evolution)

UWB (802.15)

CDMA 2000

Cassini

Mars Pathfinder & Stereo

rate 1 /

code rate [\(?\)](#)

K

constraint length [\(?\)](#)

polynomials

polynomials for the
code in decimal notation
[\(?\)](#)

Select implementation options

frame length

unpadded frame length in bits [\(?\)](#)

Vectorization level ▼

type of code [\(?\)](#)

Generate Code

Reset



“Click”: Push-button code generation

<http://www.spiral.net/software/viterbi.html>

Spiral: Generated SSE Viterbi Code

```

void viterbi_ccsds(unsigned char *Y, unsigned char *X, unsigned char *syms,
  unsigned char *dec, unsigned char *Branchtab) {
  for(int i9 = 0; i9 <= 1026; i9++) {
    unsigned char a75, a81; int a73, a92;
    ...
    a71 = ((__m128i *) X);          s18 = *(a71); a72 = (a71 + 2);
    s19 = *(a72); a73 = (4 * i9);   a74 = (syms + a73); a75 = *(a74);
    a76 = _mm_set1_epi8(a75);      a77 = ((__m128i *) Branchtab);
    a78 = *(a77);                 a79 = _mm_xor_si128(a76, a78);
    b6 = (a73 + syms);            a80 = (b6 + 1);
    a81 = *(a80);                 a82 = _mm_set1_epi8(a81);
    a83 = (a77 + 2);              a84 = *(a83);
    a85 = _mm_xor_si128(a82, a84); t13 = _mm_avg_epu8(a79, a85);
    a86 = ((__m128i ) t13);        a87 = _mm_srli_epi16(a86, 2);
    a88 = ((__m128i ) a87);
    t14 = _mm_and_si128(a88, _mm_set_epi8(63, 63, 63, 63, 63, 63, 63 , 63, 63, 63,
      63, 63, 63, 63, 63 , 63));
    t15 = _mm_subs_epu8(_mm_set_epi8(63, 63, 63, 63, 63, 63, 63 , 63, 63, 63, 63,
      63, 63, 63, 63 , 63), t14); m23 = _mm_adds_epu8(s18, t14);
    m24 = _mm_adds_epu8(s19, t15); m25 = _mm_adds_epu8(s18, t15);
    m26 = _mm_adds_epu8(s19, t14); a89 = _mm_min_epu8(m24, m23);
    ...
  }
  ...
}

```




“Click”: Push-button code generation

<http://www.spiral.net/software/viterbi.html>

Organization

- **Spiral**
- **Generating software Viterbi decoders**
- **Performance results**
- **Summary**

Organization

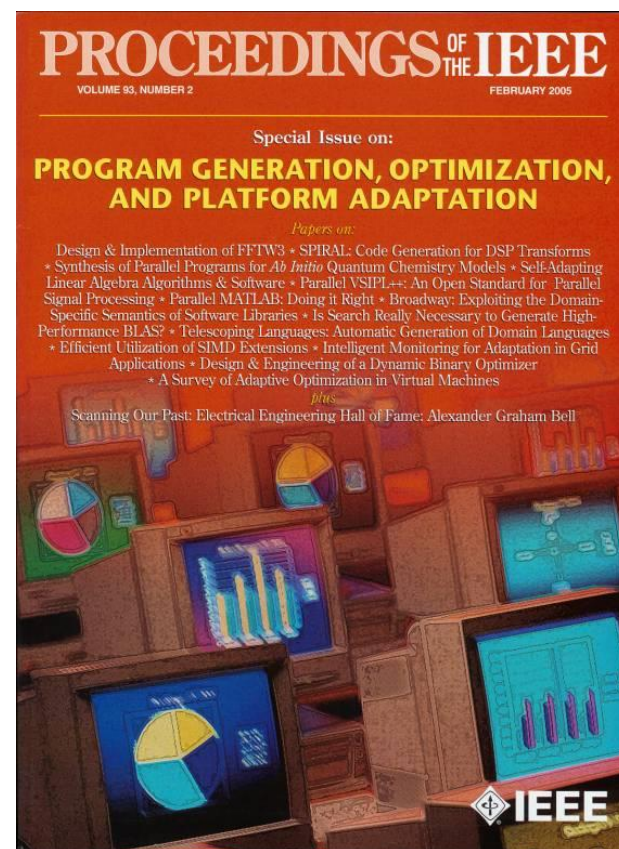
- Spiral
- Generating software Viterbi decoders
- Performance results
- Summary

Automatic Performance Tuning

- **Current vicious circle:** Whenever a new platform comes out, the same functionality needs to be rewritten and reoptimized

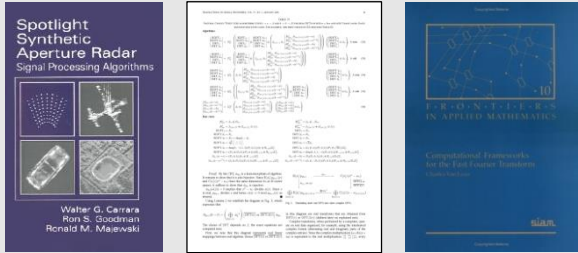
- **Automatic Performance Tuning**
 - BLAS: ATLAS, PHiPAC
 - Linear algebra: Sparsity/OSKI, Flame
 - Sorting
 - Fourier transform: FFTW
 - Linear transforms (and Viterbi): Spiral
 - ...others

New problem class: software Viterbi decoders



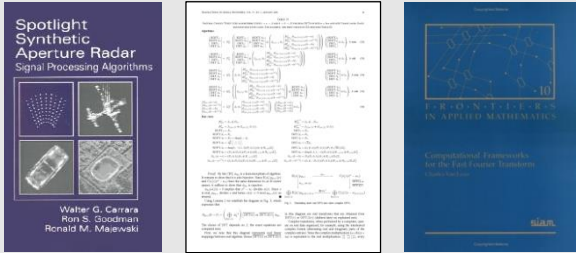
What is Spiral?

Traditionally



High performance library
optimized for given platform

Spiral Approach



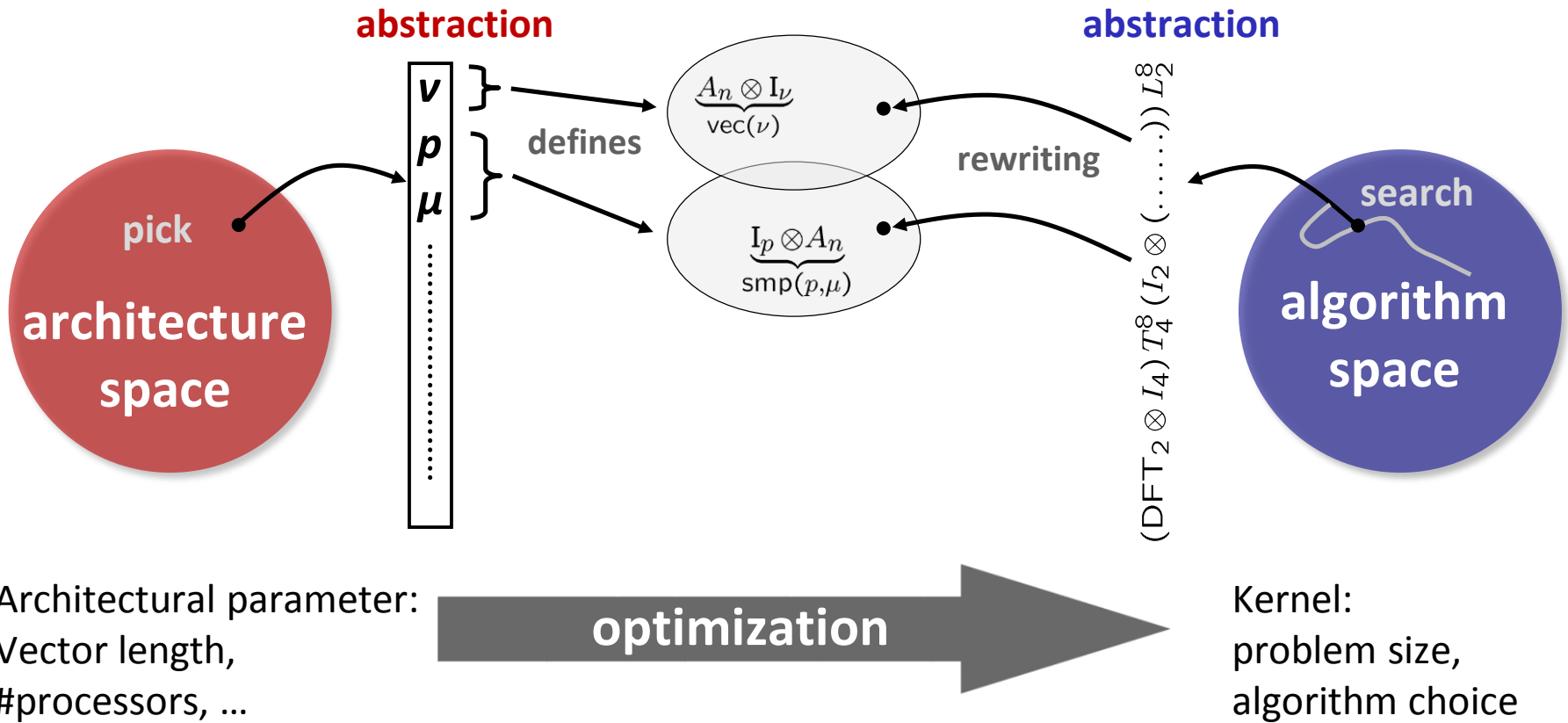
High performance library
optimized for given platform



*Comparable
performance*

Idea: Common Abstraction and Rewriting

Model: common abstraction
= spaces of matching formulas
= domain-specific language



Program Generation in Spiral

Problem specification (transform)

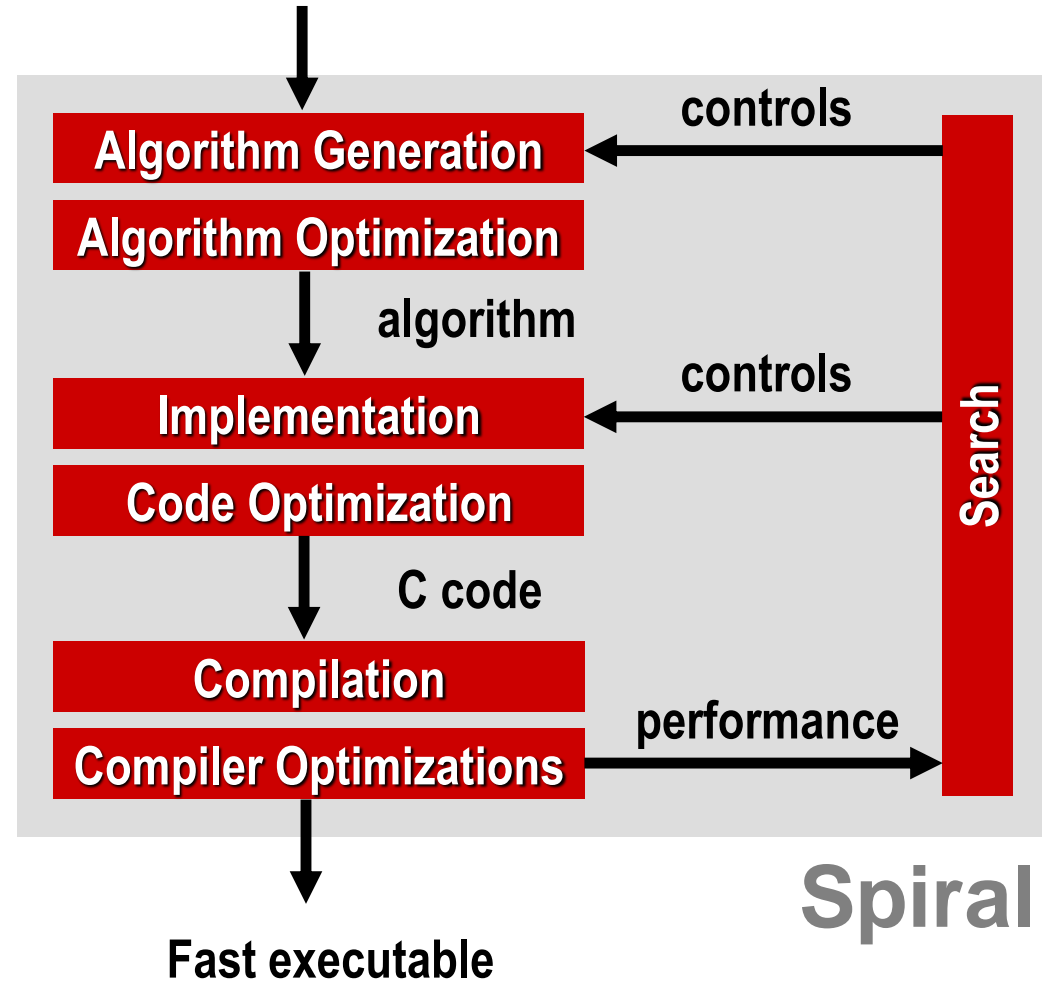
Spiral:

Complete automation of the implementation and optimization task

Basic ideas:

Declarative representation of algorithms

Rewriting systems to generate and optimize algorithms at a high level of abstraction



Some Kernels as Operator Formulas

Linear Transforms

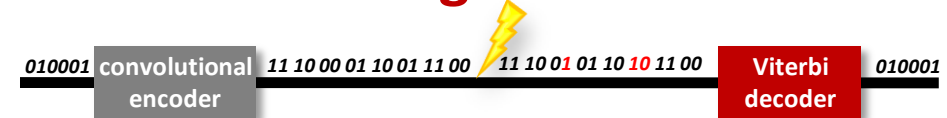
$$\begin{aligned}
 \text{DFT}_n &\rightarrow (\text{DFT}_k \otimes \text{I}_m) \text{T}_m^n (\text{I}_k \otimes \text{DFT}_m) \text{L}_k^n, \quad n = km \\
 \text{DFT}_n &\rightarrow P_n (\text{DFT}_k \otimes \text{DFT}_m) Q_n, \quad n = km, \text{ gcd}(k, m) = 1 \\
 \text{DFT}_p &\rightarrow R_p^T (\text{I}_1 \oplus \text{DFT}_{p-1}) D_p (\text{I}_1 \oplus \text{DFT}_{p-1}) R_p, \quad p \text{ prime} \\
 \text{DCT-3}_n &\rightarrow (\text{I}_m \oplus \text{J}_m) \text{L}_m^n (\text{DCT-3}_m(1/4) \oplus \text{DCT-3}_m(3/4)) \\
 &\quad \cdot (\text{F}_2 \otimes \text{I}_m) \begin{bmatrix} \text{I}_m & 0 \oplus -\text{J}_{m-1} \\ \frac{1}{\sqrt{2}}(\text{I}_1 \oplus 2\text{I}_m) \end{bmatrix}, \quad n = 2m \\
 \text{DCT-4}_n &\rightarrow S_n \text{DCT-2}_n \text{diag}_{0 \leq k < n} (1/(2 \cos((2k+1)\pi/4n))) \\
 \text{IMDCT}_{2m} &\rightarrow (\text{J}_m \oplus \text{I}_m \oplus \text{I}_m \oplus \text{J}_m) \left(\left(\begin{bmatrix} 1 \\ -1 \end{bmatrix} \otimes \text{I}_m \right) \oplus \left(\begin{bmatrix} -1 \\ -1 \end{bmatrix} \otimes \text{I}_m \right) \right) \text{J}_{2m} \text{DCT-4}_{2m} \\
 \text{WHT}_{2^k} &\rightarrow \prod_{i=1}^t (\text{I}_{2^{k_1+\dots+k_{i-1}}} \otimes \text{WHT}_{2^{k_i}} \otimes \text{I}_{2^{k_{i+1}+\dots+k_t}}), \quad k = k_1 + \dots + k_t \\
 \text{DFT}_2 &\rightarrow \text{F}_2 \\
 \text{DCT-2}_2 &\rightarrow \text{diag}(1, 1/\sqrt{2}) \text{F}_2 \\
 \text{DCT-4}_2 &\rightarrow \text{J}_2 \text{R}_{13\pi/8}
 \end{aligned}$$

Matrix-Matrix Multiplication



$$\begin{aligned}
 \text{MMM}_{1,1,1} &\rightarrow (\cdot)_1 \\
 \text{MMM}_{m,n,k} &\rightarrow (\otimes)_{m/m_b \times 1} \otimes \text{MMM}_{m_b,n,k} \\
 \text{MMM}_{m,n,k} &\rightarrow \text{MMM}_{m,n_b,k} \otimes (\otimes)_{1 \times n/n_b} \\
 \text{MMM}_{m,n,k} &\rightarrow ((\Sigma_{k/k_b} \circ (\cdot)_{k/k_b}) \otimes \text{MMM}_{m,n,k_b}) \circ \\
 &\quad ((L_{k/k_b}^{mk/k_b} \otimes \text{I}_{k_b}) \times \text{I}_{kn}) \\
 \text{MMM}_{m,n,k} &\rightarrow (L_m^{mn/n_b} \otimes \text{I}_{n_b}) \circ \\
 &\quad ((\otimes)_{1 \times n/n_b} \otimes \text{MMM}_{m,n_b,k}) \circ \\
 &\quad (\text{I}_{km} \times (L_{n/n_b}^{kn/n_b} \otimes \text{I}_{n_b}))
 \end{aligned}$$

Viterbi Decoding



$$\begin{aligned}
 \underline{\text{Vit}} &\rightarrow \left(\prod (L \times I) \circ (I \otimes C) \right) \circ \text{Id} \\
 \text{vec}(v) &\quad \underbrace{\hspace{10em}}_{\text{vec}(v)} \\
 &\rightarrow \left(\prod \underbrace{(L \times I) \circ (I \otimes C)}_{\text{vec}(v)} \right) \circ \text{Id} \\
 \mathcal{L} &\rightarrow \left(\prod (L \otimes \text{I}_v \times I) \circ (I \otimes C \otimes \text{I}_v) \circ (\vec{L} \times I) \right) \circ \text{Id} \\
 &\rightarrow \prod (L \otimes \text{I}_v \times I) \circ (I \otimes (B \otimes \text{I}_v)) \circ (\vec{L} \times I)
 \end{aligned}$$

Synthetic Aperture Radar (SAR)



$$\begin{aligned}
 \text{SAR}_{k \times m \rightarrow n \times n} &\rightarrow \text{DFT}_{n \times n} \circ \text{Interp}_{k \times m \rightarrow n \times n} \\
 \text{DFT}_{n \times n} &\rightarrow (\text{DFT}_n \otimes \text{I}_n) \circ (\text{I}_n \otimes \text{DFT}_n) \\
 \text{Interp}_{k \times m \rightarrow n \times n} &\rightarrow (\text{Interp}_{k \rightarrow n} \otimes_i \text{I}_n) \circ (\text{I}_k \otimes_i \text{Interp}_{m \rightarrow n}) \\
 \text{Interp}_{r \rightarrow s} &\rightarrow \left(\bigoplus_{i=0}^{n-2} \text{InterpSeg}_k \right) \oplus \text{InterpSegPruned}_{k,\ell} \\
 \text{InterpSeg}_k &\rightarrow G_f^{u,n \rightarrow k} \circ \text{iPrunedDFT}_{n \rightarrow u,n} \circ \left(\frac{1}{n} \right) \circ \text{DFT}_n
 \end{aligned}$$

Same Approach for Different Paradigms

Threading:

$$\begin{aligned}
 \underbrace{\text{DFT}_{mn}}_{\text{smp}(p,\mu)} &\rightarrow \underbrace{\left((\text{DFT}_m \otimes \text{I}_n) \text{T}_n^{mn} (\text{I}_m \otimes \text{DFT}_n) \text{L}_m^{mn} \right)}_{\text{smp}(p,\mu)} \\
 &\dots \\
 &\rightarrow \underbrace{\left(\text{DFT}_m \otimes \text{I}_n \right)}_{\text{smp}(p,\mu)} \underbrace{\text{T}_n^{mn}}_{\text{smp}(p,\mu)} \underbrace{\left(\text{I}_m \otimes \text{DFT}_n \right)}_{\text{smp}(p,\mu)} \underbrace{\text{L}_m^{mn}}_{\text{smp}(p,\mu)} \\
 &\dots \\
 &\rightarrow \left(\text{L}_m^{mp} \otimes \text{I}_{n/p\mu} \otimes \mu \text{I}_\mu \right) \left(\text{I}_p \otimes \parallel (\text{DFT}_m \otimes \text{I}_{n/p}) \right) \left(\text{L}_p^{mp} \otimes \text{I}_{n/p\mu} \otimes \mu \text{I}_\mu \right) \\
 &\quad \left(\bigoplus_{i=0}^{p-1} \text{T}_n^{mn,i} \right) \left(\text{I}_p \otimes \parallel (\text{I}_{m/p} \otimes \text{DFT}_n) \right) \left(\text{I}_p \otimes \parallel \text{L}_{m/p}^{mn/p} \right) \left(\text{L}_p^{pn} \otimes \text{I}_{m/p\mu} \otimes \mu \text{I}_\mu \right)
 \end{aligned}$$

Vectorization:

$$\begin{aligned}
 \underbrace{\left(\text{DFT}_{mn} \right)}_{\text{vec}(\nu)} &\rightarrow \underbrace{\left((\text{DFT}_m \otimes \text{I}_n) \text{T}_n^{mn} (\text{I}_m \otimes \text{DFT}_n) \text{L}_m^{mn} \right)}_{\text{vec}(\nu)} \\
 &\dots \\
 &\rightarrow \underbrace{\left(\text{DFT}_m \otimes \text{I}_n \right)^\nu}_{\text{vec}(\nu)} \underbrace{\left(\text{T}_n^{mn} \right)^\nu}_{\text{vec}(\nu)} \underbrace{\left(\text{I}_m \otimes \text{DFT}_n \right)^\nu}_{\text{vec}(\nu)} \underbrace{\text{L}_m^{mn}^\nu}_{\text{vec}(\nu)} \\
 &\dots \\
 &\rightarrow \left(\text{I}_{mn/\nu} \otimes \underbrace{\text{L}_\nu^{2\nu}}_{\text{sse}} \right) \underbrace{\left(\text{DFT}_m \otimes \text{I}_{n/\nu} \right)^\nu}_{\text{sse}} \underbrace{\left(\text{T}_n^{mn} \right)^\nu}_{\text{sse}} \\
 &\quad \left(\text{I}_{m/\nu} \otimes \left(\text{L}_\nu^{n/\nu} \vec{\otimes} \text{I}_\nu \right) \right) \left(\text{I}_{n/\nu} \otimes \left(\text{L}_\nu^{2\nu} \vec{\otimes} \text{I}_\nu \right) \right) \left(\text{I}_2 \otimes \underbrace{\text{L}_\nu^{\nu^2}}_{\text{sse}} \right) \left(\text{L}_2^{2\nu} \vec{\otimes} \text{I}_\nu \right) \left(\text{DFT}_n \vec{\otimes} \text{I}_\nu \right) \\
 &\quad \left(\text{L}_m^{mn} \otimes \text{I}_2 \right) \vec{\otimes} \text{I}_\nu \left(\text{I}_{mn/\nu} \otimes \underbrace{\text{L}_2^{2\nu}}_{\text{sse}} \right)
 \end{aligned}$$

GPUs:

$$\begin{aligned}
 \underbrace{\left(\text{DFT}_{r^k} \right)}_{\text{gpu}(t,c)} &\rightarrow \underbrace{\left(\prod_{i=0}^{k-1} \text{L}_r^{r^k} \left(\text{I}_{r^{k-1}} \otimes \text{DFT}_r \right) \left(\text{L}_{r^{k-i-1}}^{r^k} \left(\text{I}_{r^i} \otimes \text{T}_{r^{k-i-1}}^{r^{k-i}} \right) \text{L}_{r^{i+1}}^{r^k} \right) \right)}_{\text{gpu}(t,c)} \text{R}_r^{r^k} \\
 &\dots \\
 &\rightarrow \left(\prod_{i=0}^{k-1} \left(\text{L}_r^{r^n/2} \vec{\otimes} \text{I}_2 \right) \left(\text{I}_{r^{n-1}/2} \otimes \times \underbrace{\left(\text{DFT}_r \vec{\otimes} \text{I}_2 \right) \text{L}_r^{2r}}_{\text{shd}(t,c)} \right) \text{T}_i \right) \\
 &\quad \left(\text{L}_r^{r^n/2} \vec{\otimes} \text{I}_2 \right) \left(\text{I}_{r^{n-1}/2} \otimes \times \underbrace{\text{L}_r^{2r}}_{\text{shd}(t,c)} \right) \left(\text{R}_r^{r^{n-1}} \vec{\otimes} \text{I}_r \right)
 \end{aligned}$$

Verilog for FPGAs:

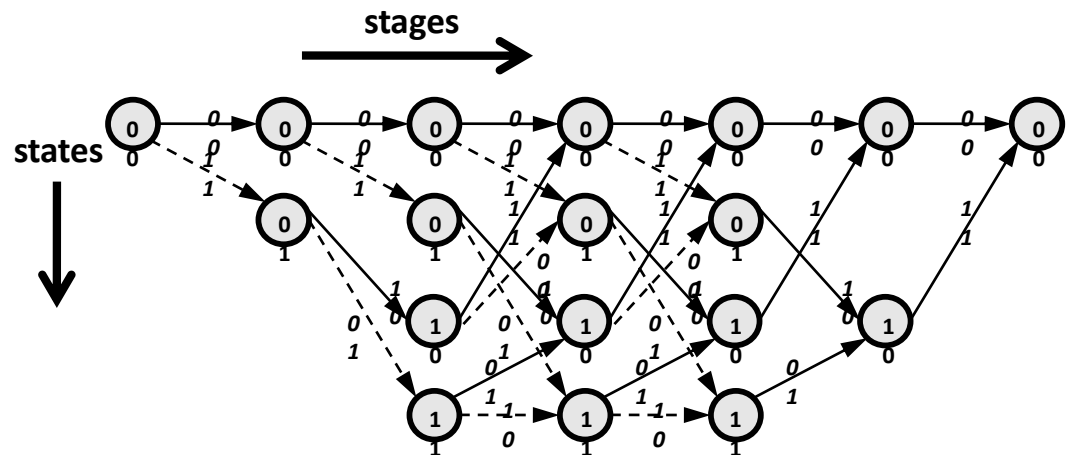
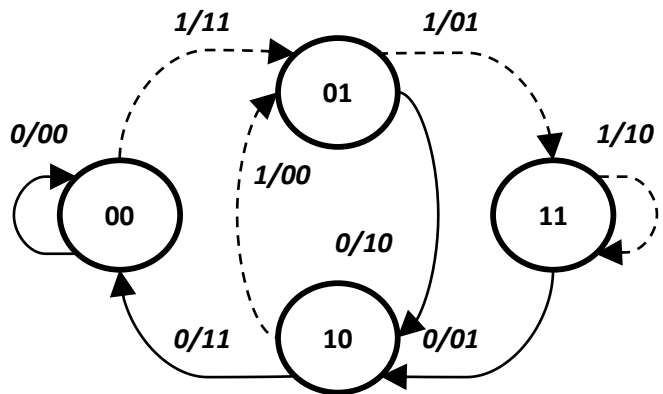
$$\begin{aligned}
 \underbrace{\left(\text{DFT}_{r^k} \right)}_{\text{stream}(r^s)} &\rightarrow \underbrace{\left[\prod_{i=0}^{k-1} \text{L}_r^{r^k} \left(\text{I}_{r^{k-1}} \otimes \text{DFT}_r \right) \left(\text{L}_{r^{k-i-1}}^{r^k} \left(\text{I}_{r^i} \otimes \text{T}_{r^{k-i-1}}^{r^{k-i}} \right) \text{L}_{r^{i+1}}^{r^k} \right) \right]}_{\text{stream}(r^s)} \text{R}_r^{r^k} \\
 &\dots \\
 &\rightarrow \underbrace{\left[\prod_{i=0}^{k-1} \underbrace{\text{L}_r^{r^k}}_{\text{stream}(r^s)} \left(\text{I}_{r^{k-1}} \otimes \text{DFT}_r \right) \left(\text{L}_{r^{k-i-1}}^{r^k} \left(\text{I}_{r^i} \otimes \text{T}_{r^{k-i-1}}^{r^{k-i}} \right) \text{L}_{r^{i+1}}^{r^k} \right) \right]}_{\text{stream}(r^s)} \text{R}_r^{r^k} \\
 &\dots \\
 &\rightarrow \underbrace{\left[\prod_{i=0}^{k-1} \underbrace{\text{L}_r^{r^k}}_{\text{stream}(r^s)} \left(\text{I}_{r^{k-s-1}} \otimes_s \left(\text{I}_{r^{s-1}} \otimes \text{DFT}_r \right) \right) \text{T}'_i \right]}_{\text{stream}(r^s)} \text{R}_r^{r^k}
 \end{aligned}$$

Organization

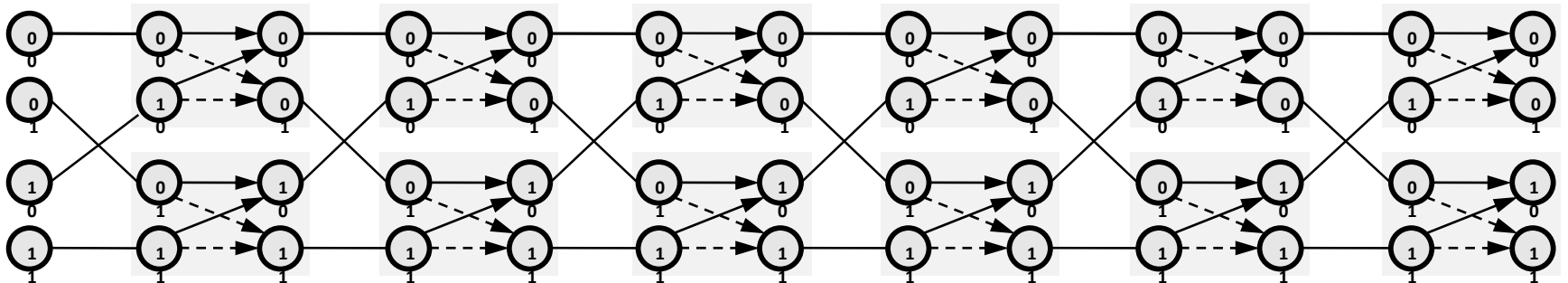
- Spiral
- **Generating software Viterbi decoders**
- Performance results
- Summary

Structure of Viterbi Decoders

State machine



Viterbi trellis (data flow)



Key observation: similarity to Walsh-Hadamard transform (WHT)

Viterbi Language (VL)

VL in Backus-Naur Form (BNF)

$\langle \text{op} \rangle ::=$	$F_{K,F}$	<i>Viterbi forward pass</i>
	$ I_n$	<i>identity</i>
	$ L_n^{mn}$	<i>stride permutation</i>
	$ B_{i,j}$	<i>Viterbi butterfly</i>
	$ \langle \text{op} \rangle \langle \text{op} \rangle$	<i>composition</i>
	$ \Pi \langle \text{op} \rangle$	<i>iterative composition</i>
	$ \langle \text{op} \rangle \otimes \langle \text{op} \rangle$	<i>tensor product</i>

Viterbi decoder forward pass in VL

$$F_{K,F} \rightarrow \prod_{i=1}^F \left((I_{2^{K-2}} \otimes_j B_{F-i,j}) L_{2^{K-2}}^{2^{K-1}} \right)$$

$$B_{i,j} : \begin{cases} \pi_U = \min_{d_U} (\pi_A + \beta_{A \rightarrow U}, \pi_B + \beta_{B \rightarrow U}) \\ \pi_V = \min_{d_V} (\pi_A + \beta_{A \rightarrow V}, \pi_B + \beta_{B \rightarrow V}) \end{cases}$$

Compiling VL To Code

construct	code
$y = (CD)x$	<pre>t = D(x); y = C(t);</pre>
$y = \prod_{i=0}^{l-1} C^i x$	<pre>y = C(l-1, x); for (i=l-2;i>=0;i--) y = C(i, y);</pre>
$y = (I_m \otimes_j C_n^j)x$	<pre>for (j=0;j<m;j++) y[j*n:j*n+n-1] = C(j, x[j*n:j*n+n-1]);</pre>
$y = L_m^{mn} x$	<pre>for (i=0;i<m;i++) for (j=0;j<n;j++) y[i+m*j]=x[n*i+j];</pre>
$y = B_{i,j} x$	<i>see equation last slide</i>

Vectorization Through Rewriting

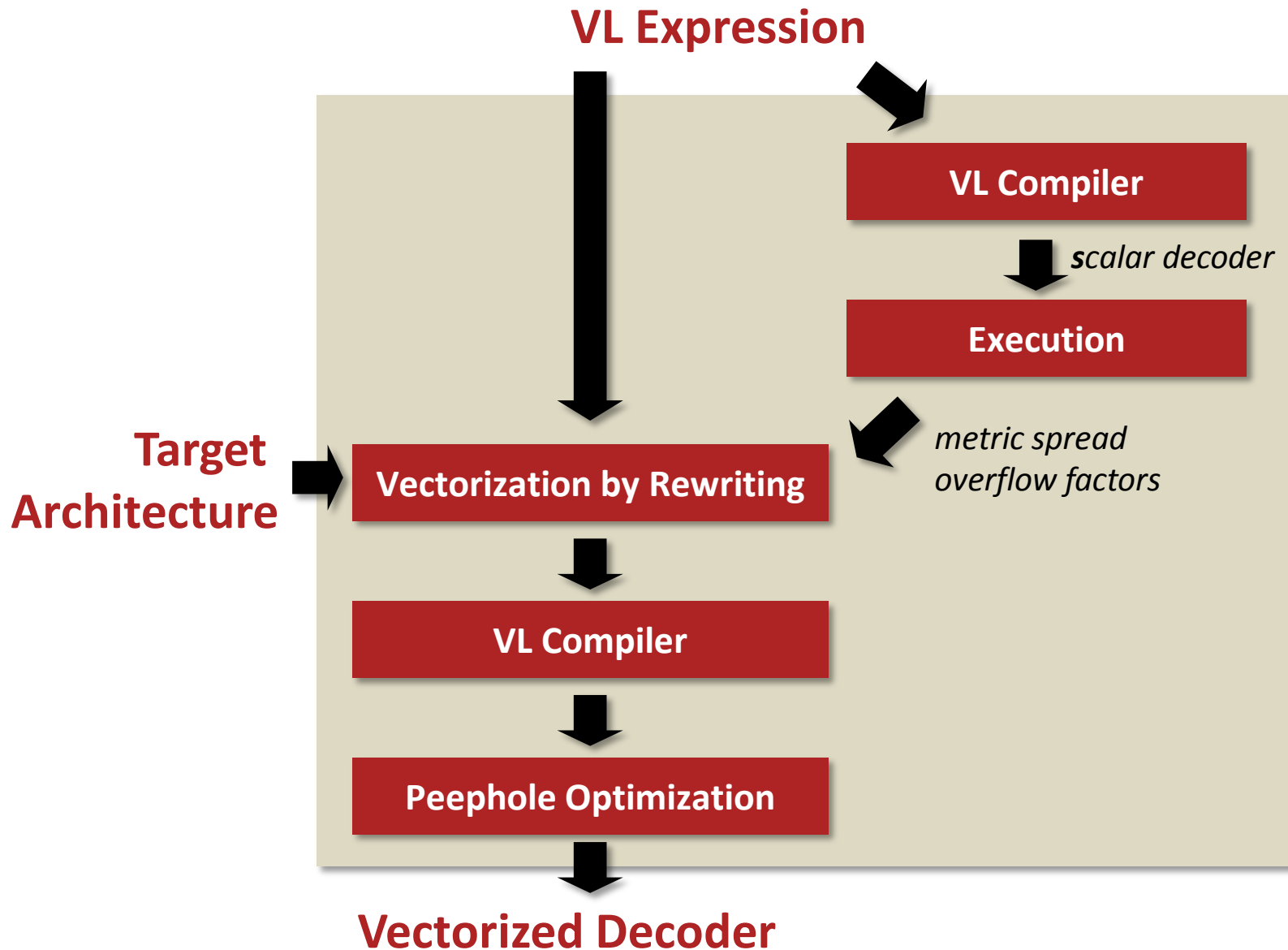
Vectorization Rule Set

$$\begin{aligned}
 \underline{CD} \nu &\rightarrow \underline{C} \nu \underline{D} \nu \\
 \underline{\prod C} \nu &\rightarrow \prod \underline{C} \nu \\
 \underline{I_m \otimes_j C^j} \nu &\rightarrow I_m \otimes_j \underline{C^j} \nu \\
 \underline{C \otimes I_\nu} \nu &\rightarrow C \bar{\otimes} I_\nu \\
 \underline{B_{i,l\nu+j} \otimes_j I_\nu} \nu &\rightarrow \vec{B}_{i,l}^\nu \\
 \underline{L_\nu^{2\nu}} \nu &\rightarrow \vec{L}_\nu^{2\nu}
 \end{aligned}$$

Vectorized Viterbi Decoder

$$\underline{F_{K,F}} \nu \rightarrow \prod_{i=1}^F \left(\left(I_{2^{K-2}/\nu} \otimes_{j_1} \vec{L}_\nu^{2\nu} \vec{B}_{F-i,j_1}^\nu \right) \left(L_{2^{K-2}/\nu}^{2^{K-1}/\nu} \bar{\otimes} I_\nu \right) \right)$$

VL Compilation System



Organization

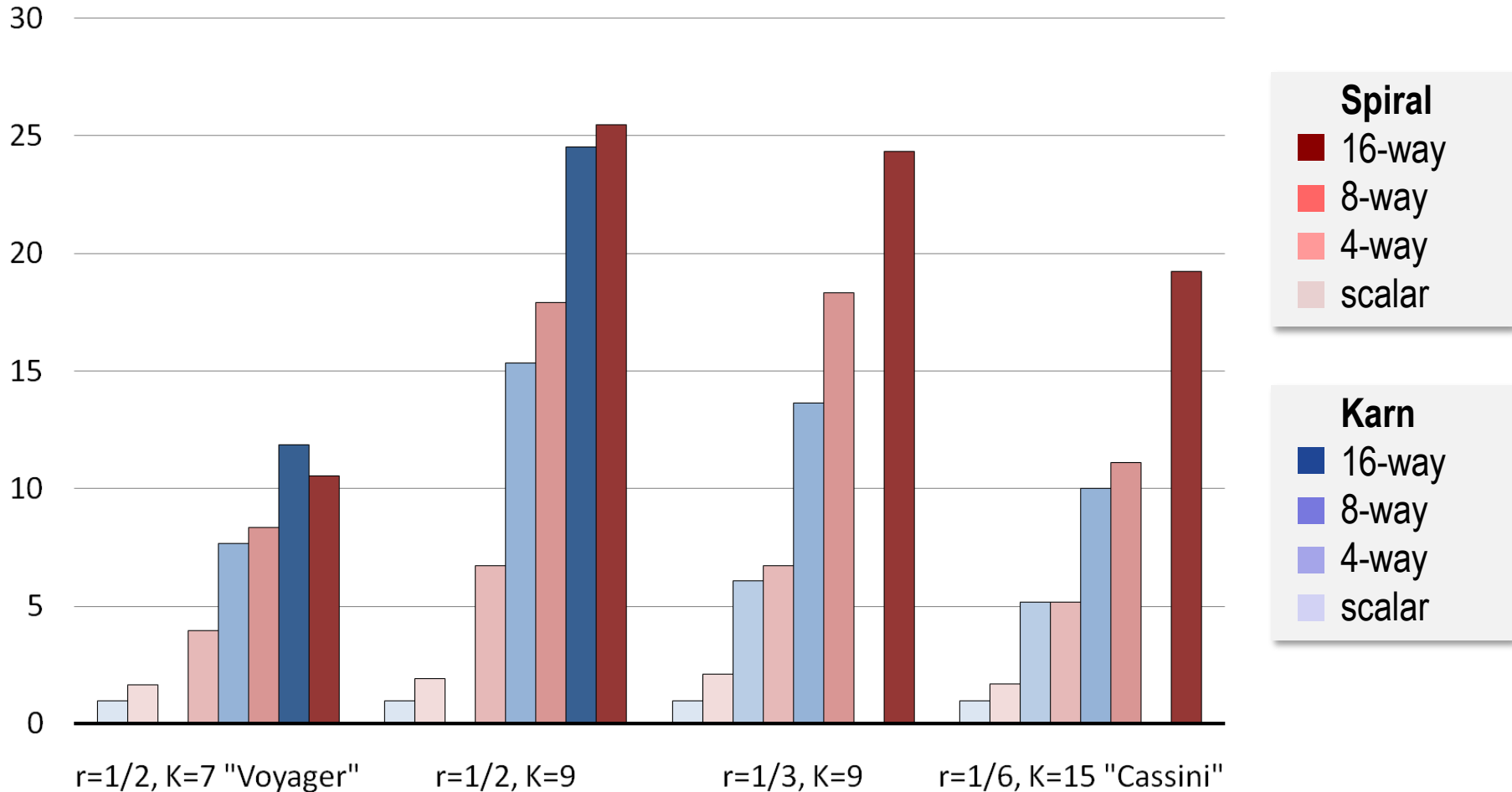
- Spiral
- Generating Software Viterbi Decoders
- Performance results
- Summary

Comparison to Hand-Tuned Code

Karn's implementation: hand-written assembly for 4 specific Viterbi codes

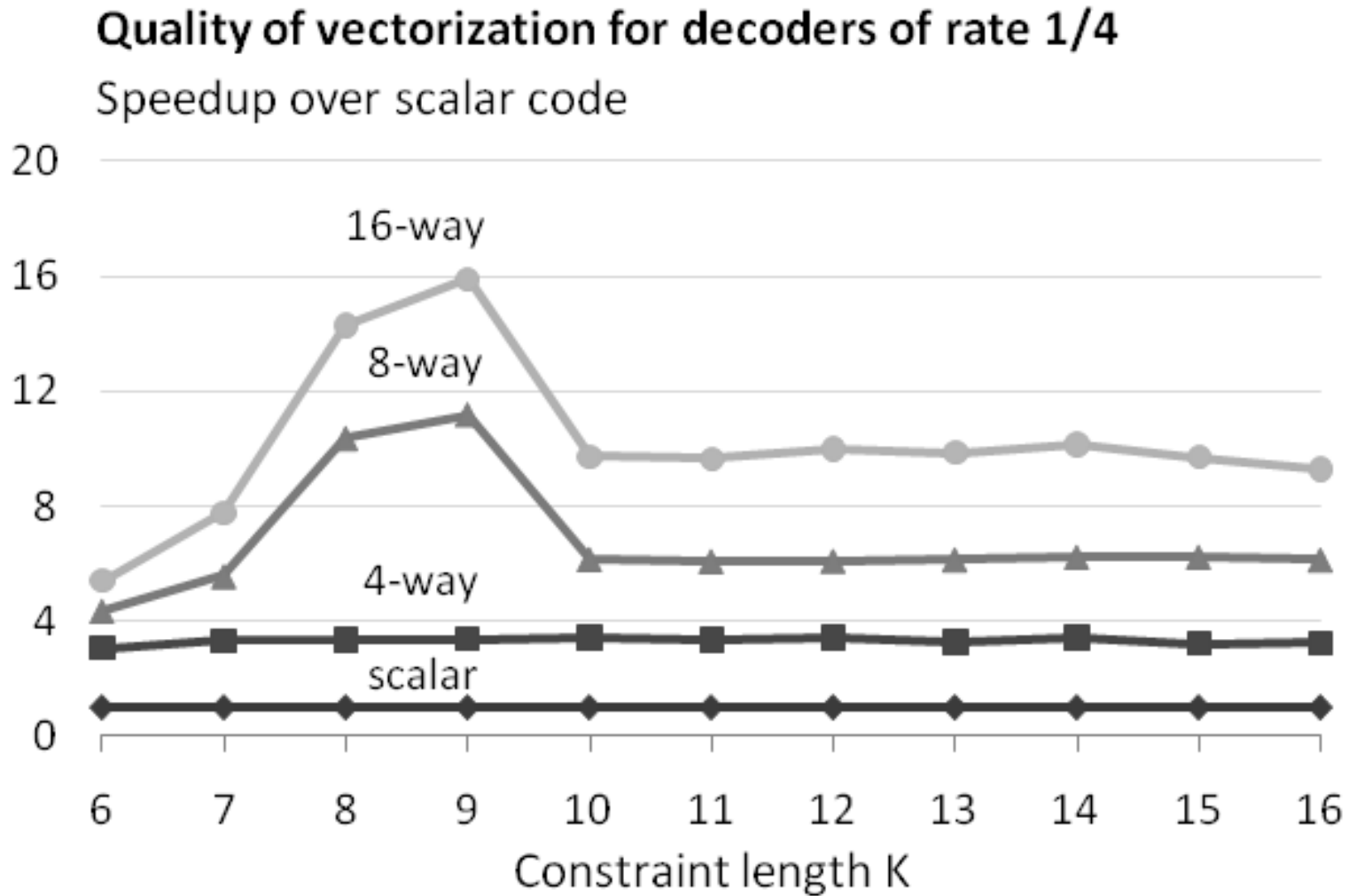
Performance Gain of Various Generated Viterbi Decoders

Speedup over Karn's C implementation



Single core of Core2 Extreme (quad-core), 3 GHz, Intel C++ compiler 10.0

Vectorization Speed-Up

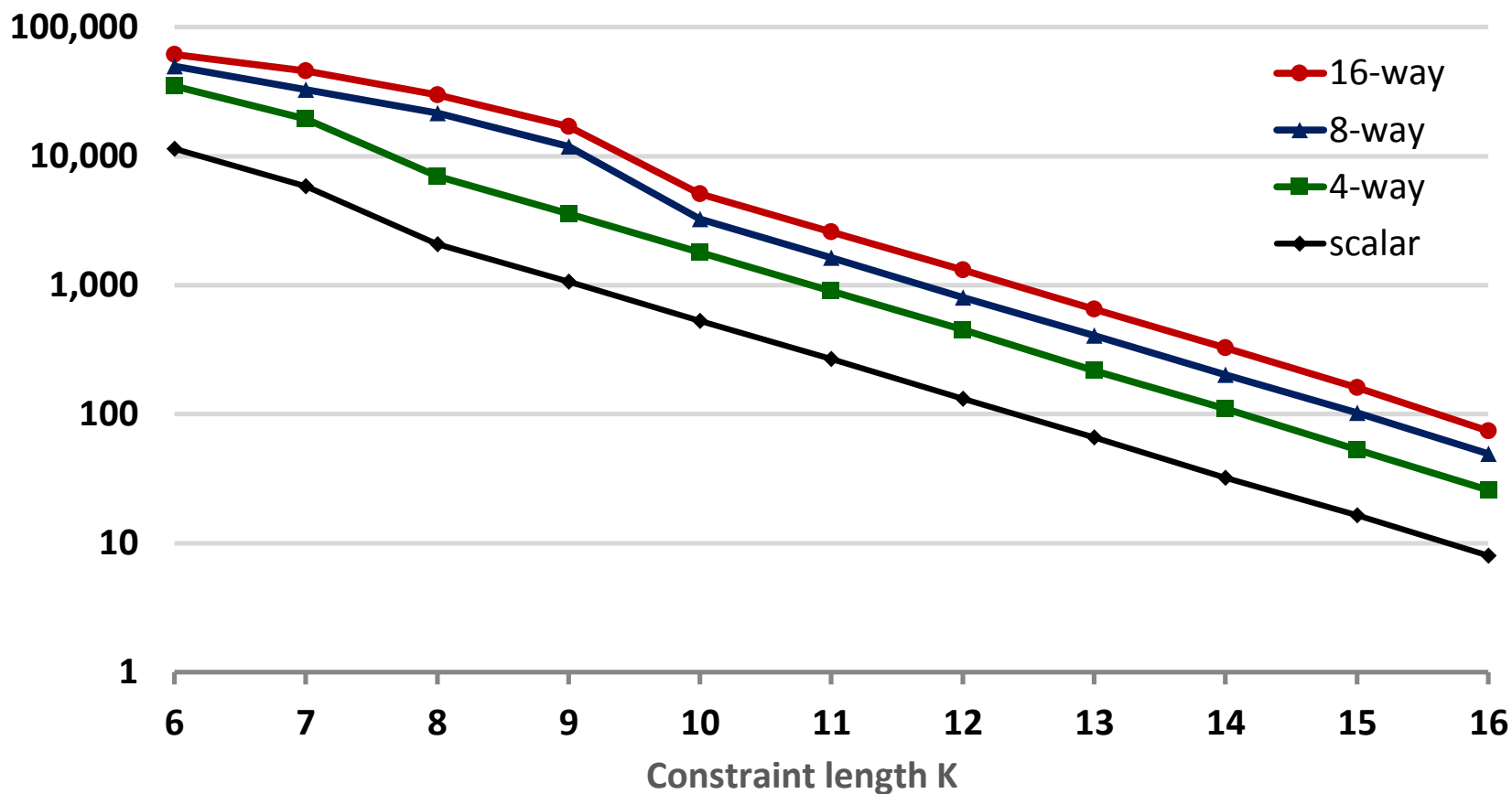


Single core of Core2 Extreme (quad-core), 3 GHz, Intel C++ compiler 10.0

Data Rate Results

Decoders for rate 1/4

Performance (kbit/s)



Single core of Core2 Extreme (quad-core), 3 GHz, Intel C++ compiler 10.0

Organization

- Spiral
- Generating Software Viterbi Decoders
- Performance results
- **Summary**

Summary

- Platforms are powerful yet complicated optimization will stay a hard problem

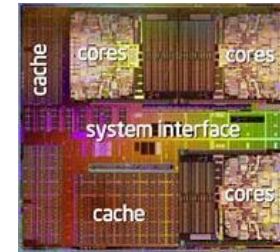
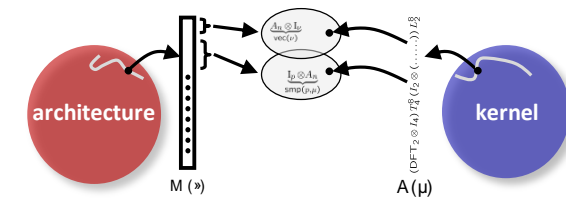


Image: Intel

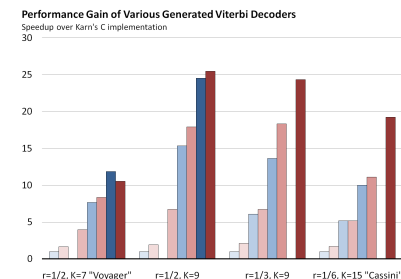
- Automatic generation of Viterbi decoder from high-level specification

$$F_{K,F} \rightarrow \prod_{i=1}^F \left((I_{2^{K-2}} \otimes_j B_{F-i,j}) L_{2^{K-2}}^{2^{K-1}} \right)$$

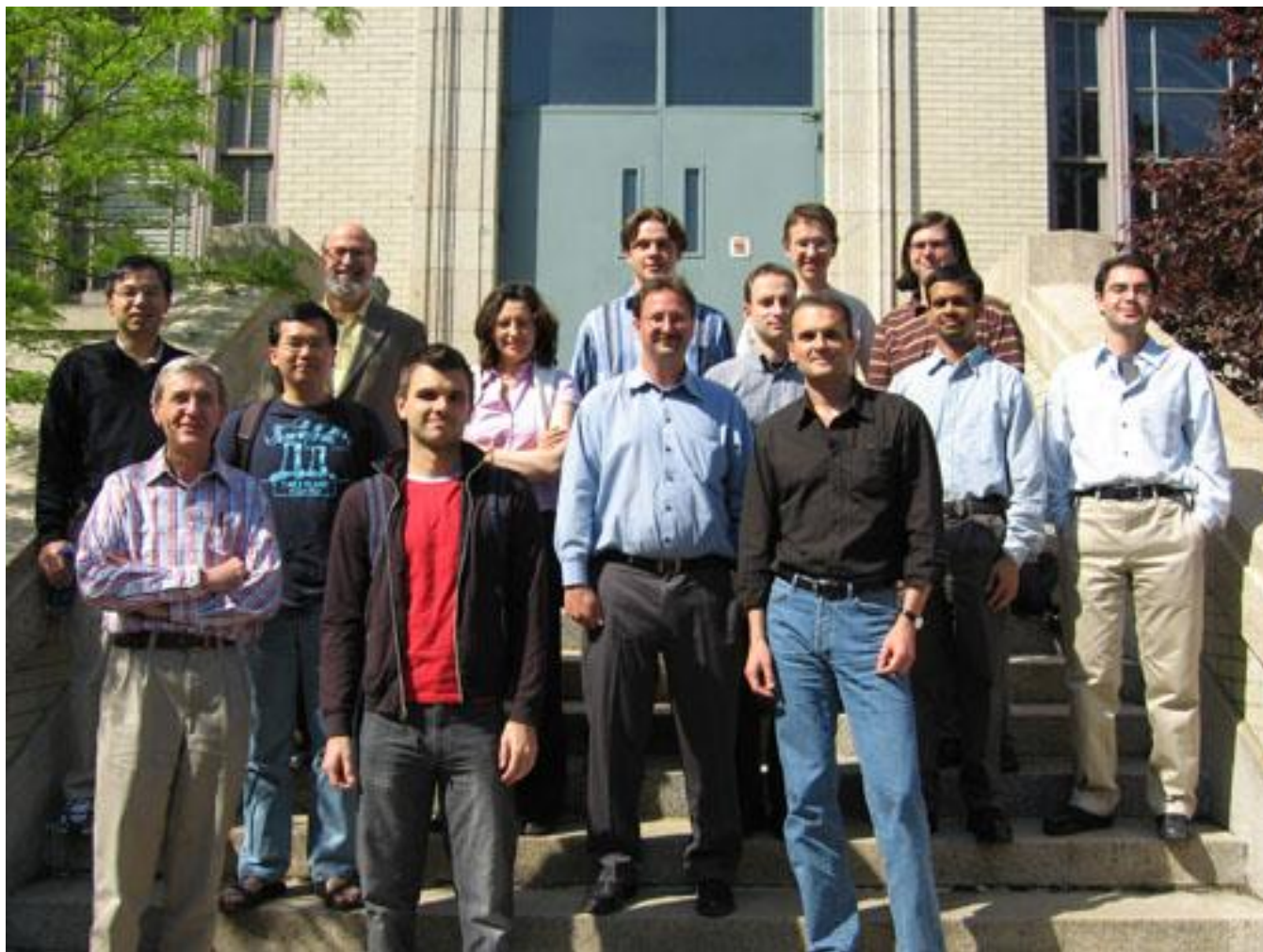
- Spiral: program generation and autotuning can provide full automation



- Performance of Spiral's Viterbi decoders is competitive with expert hand tuning



(Part of the) Spiral Team



www.spiral.net