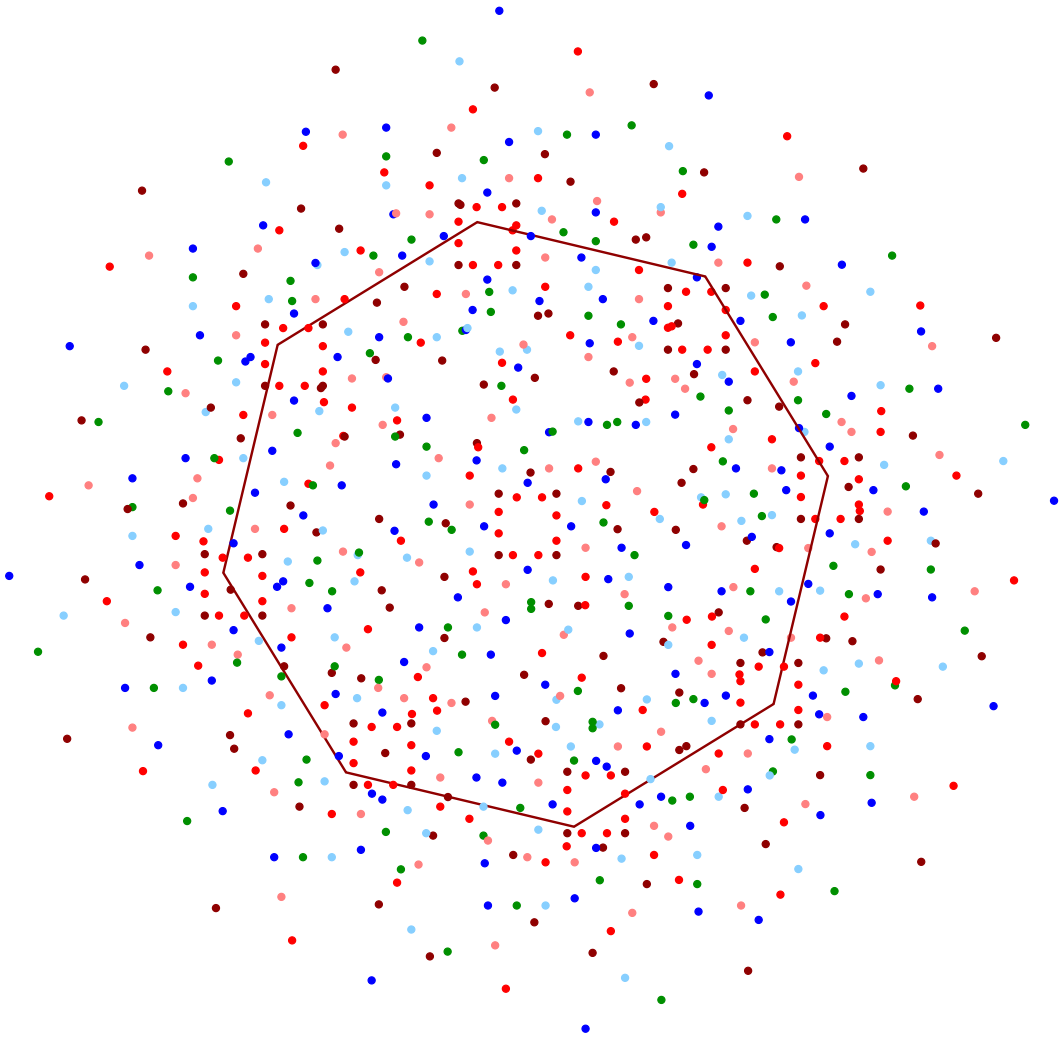
# Approximation of Complex Numbers by Cyclotomic Integers



Amin Shokrollahi

Joint work with Joe Buhler and Volker Stemann

# An Approximation Problem

Given a real number $\alpha$ between 0 and 1 and an integer $M$, approximate $\alpha$ by $a + b\sqrt{2}$ for integers $a$ and $b$ such that $|a|, |b| \leq M$.

The problem arises in connection with symbolic computation.

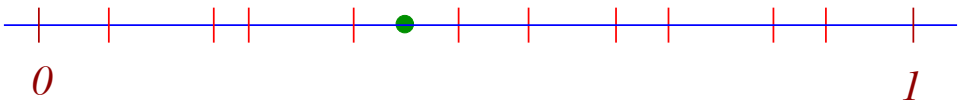
Instead of performing operations on floating points, we want to perform them on vectors of integers. (Exact arithmetic.)

An example will be given later.

# Best approximation error

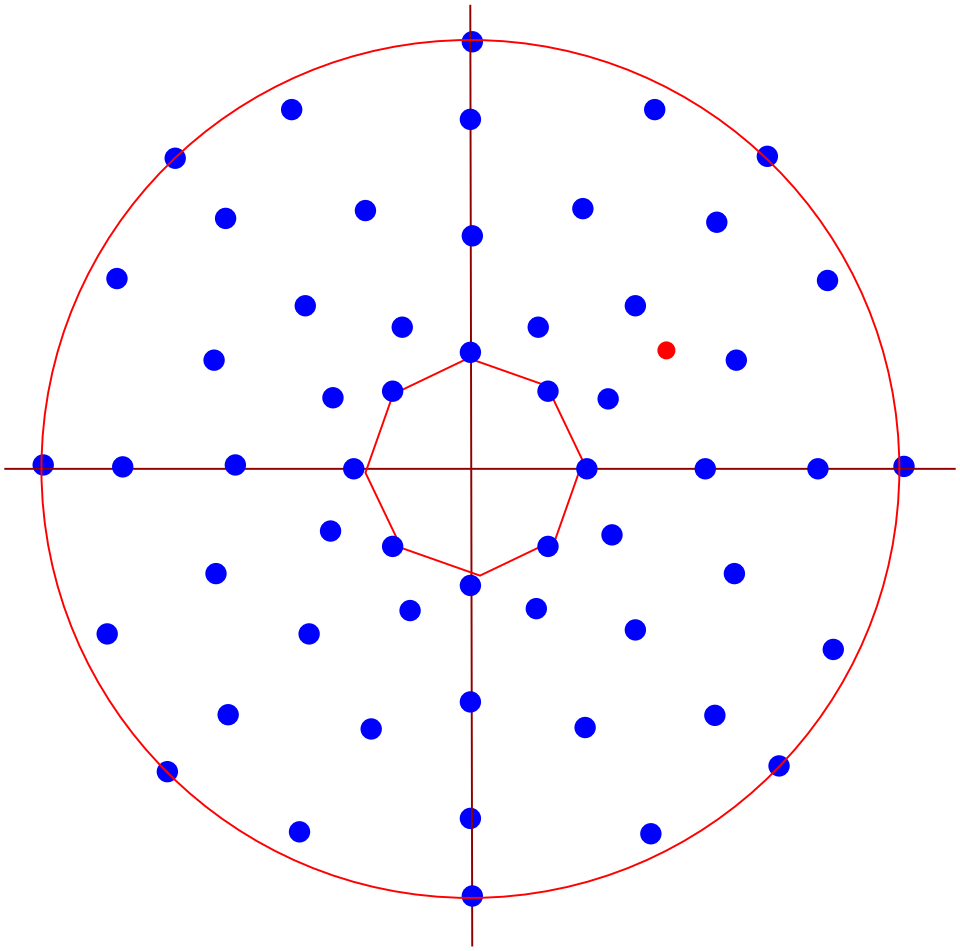
Given $a$, $|a| \leq M$, there is at most one $b$ with $|b| \leq M$ and $0 \leq a + b\sqrt{2} \leq 1$.

Consequence: there are $\alpha$'s that cannot be approximated with an error less than $1/2M$.



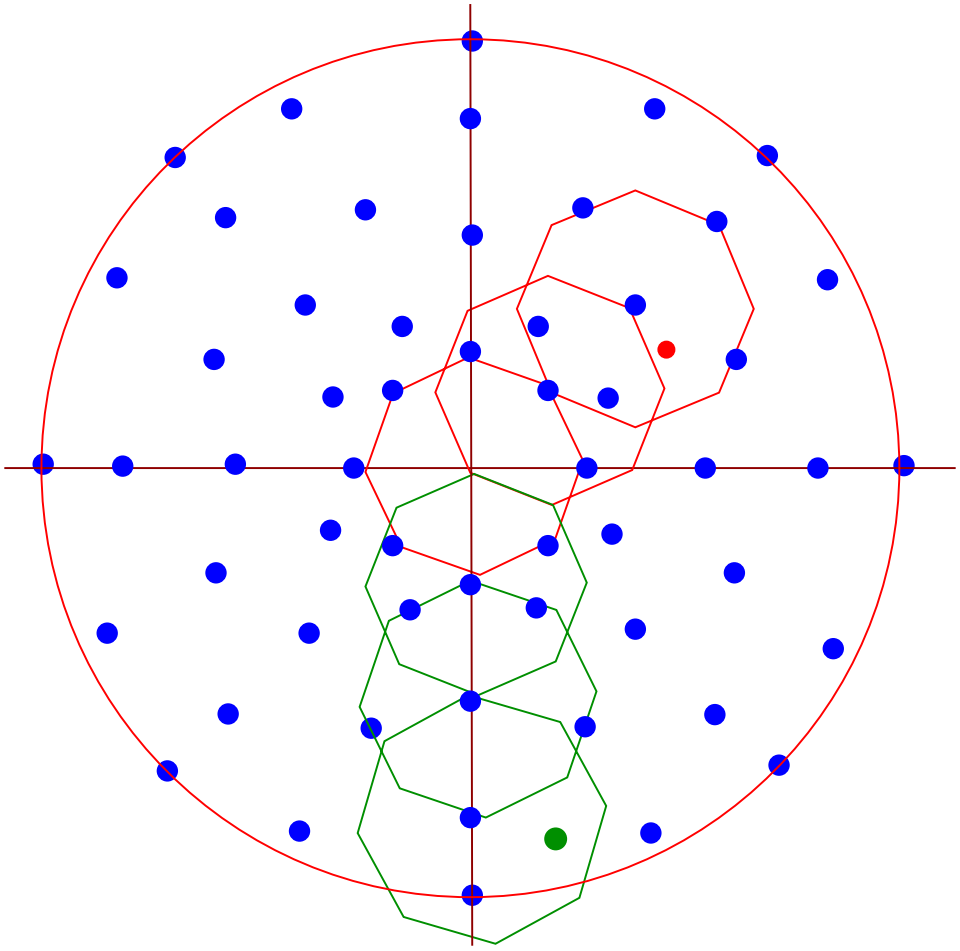
# Related Complex Problem

Given a complex number $z$ inside the unit circle, and an integer $M$, approximate $z$ by an expression of the form $a + b\zeta + c\zeta^2 + d\zeta^4$, where $\zeta = \exp(2\pi i/8)$ and $a, b, c, d$ are integers of absolute value $\leq M$.



# Games' Algorithm – 1986

Games proposed in 1986 the following algorithm for approximation.



Running time: $O(M)$.

Approximation error: $O(1/M)$.

disadvantage: Complicated search structures, not suited for real time applications.

# A Simple Algorithm

Suppose we want to approximate 0.1, with $M = 64$.

Let $E := \{-41 + 29\sqrt{2}, 17 - 12\sqrt{2}\} =: \{\varepsilon_1, \varepsilon_2\}$.

Start with the approximation $a_1 := 0$:

$a_2 := a_1 + \varepsilon_2 = 17 - 12\sqrt{2} \sim 0.0293$

$a_3 := a_2 + \varepsilon_1 = -24 + 17\sqrt{2} \sim 0.0416$

$a_4 := a_3 + \varepsilon_2 = -7 + 5\sqrt{2} \sim 0.0711$

$a_5 := a_4 + \varepsilon_2 = 10 - 7\sqrt{2} \sim 0.1005$

$a_6 := a_5 + \varepsilon_1 = -31 + 22\sqrt{2} \sim 0.1126$.

Hence, we stop with the approximation $a_5 = 10 - 7\sqrt{2}$.

The approximation error is 0.005...

Where does $E$ come from?

# Continued Fractions

We need to construct a set $E$ consisting of two small positive $M$-bounded elements of different signature.

We use the continued fraction expansion of $\sqrt{2}$:

$$\sqrt{2} = 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{\cdots}}}.$$

This gives the sequence of convergents

$$\sqrt{2} \sim \frac{1}{1}, \frac{3}{2}, \frac{7}{5}, \ldots, \frac{P_\ell}{Q_\ell}, \ldots$$

where $P_\ell = 2P_{\ell-1} + P_{\ell-2}$ and $Q_\ell = 2Q_{\ell-1} + Q_{\ell-2}$.

One has $(-1)^\ell(P_\ell - Q_\ell\sqrt{2}) = (-1+\sqrt{2})^\ell > 0$. Hence we can set

$$E := \left\{(-1)^\ell(P_\ell - Q_\ell\sqrt{2}), (-1)^{\ell+1}(P_{\ell+1} - Q_{\ell+1}\sqrt{2})\right\}$$

for suitable $\ell$.

The final result is an approximation algorithm with worst case error of $1.71/M$ which compares very well with the lower bound $\Omega(1/M)$.

The algorithm can be modified to run in time $O(\log(M))$.

# Approximation

For $\zeta := e^{2\pi i/2^n}$ let $\mathbb{Z}[\zeta]_M$ be the set of integral linear combinations of powers of $\zeta$ with coefficients bounded by $M$ in absolute value.

Design an algorithm that approximates a given complex number inside the unit circle by an element of the set $\mathbb{Z}[\zeta]_M$.

- Cozzens and Finkelstein'85: general algorithm, optimal error, infeasible since exhaustive search.

- Games'86: special case $n = 3$, optimal approximation error $\sim 1/M$, running time $O(M)$, however: complicated search structures, not suited for real time applications.

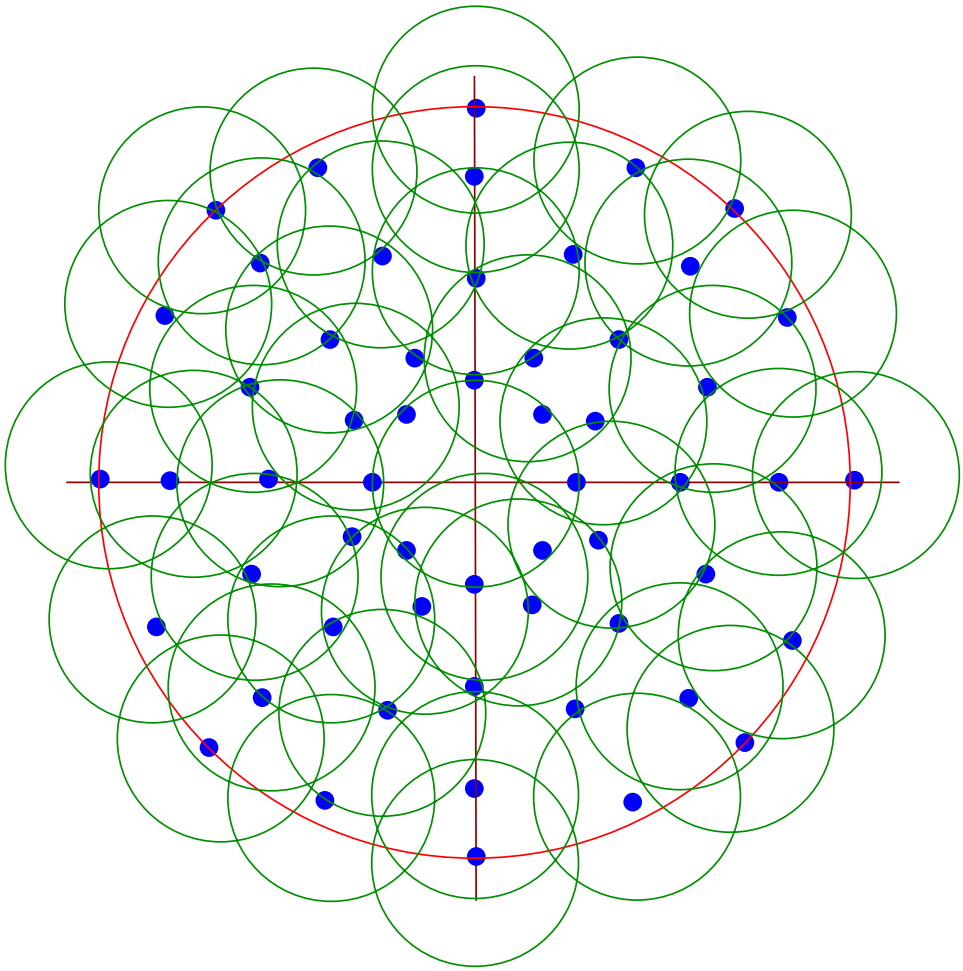
Our algorithm: general, close to optimal error, suited for real time applications.

Example: $n = 3 \rightarrow$ approximation error $\sim 1/M$, running time $O(\log(M))$.

Practical setting: $n = 4$, approximation error $\sim 1/M^3$, running time $O(\log(M))$.

# What can we expect?

For fixed $n$ any approximation algorithm has a worst case error of $\Omega(1/M^{2^{n-2}-1})$.



It is sufficient to approximate real numbers between 0 and 1 by real elements of $\mathbb{Z}[\zeta]_M$, i.e., by elements of the form

$$a_0 + a_1 2\cos\frac{2\pi}{2^n} + \ldots + a_{2^{n-2}-1} 2\cos\frac{2\cdot(2^{n-2}-1)\pi}{2^n}$$

where $|a_0|, \ldots, |a_{2^{n-2}-1}| \le M$.

# 16th Roots of Unity

Example: Approximation of $0.1$ in $\mathbb{Z}[\zeta]_{10}$

Let $\theta_0 := 1$, $\theta_1 := \sqrt{2+\sqrt{2}}$, $\theta_2 := \sqrt{2}$, $\theta_3 := \sqrt{2-\sqrt{2}}$. We use a set $E$ whose elements have the following representation with respect to the above basis:

$$\begin{aligned} E &:= \{(-3,1,3,-4),(6,-3,-3,5),(-5,5,-4,2),\\ &\quad (10,-9,7,-4),(-5,-2,4,4),(8,2,-5,-6)\}\\ &=: \{\varepsilon_1,\ldots,\varepsilon_6\}. \end{aligned}$$

We start with $a_1 := 0$:

$a_2 := a_1 + \varepsilon_1 = (-3,1,3,-4) \sim 0.0289$

$a_3 := a_2 + \varepsilon_2 = (3,-2,0,1) \sim 0.0698$

$a_4 := a_3 + \varepsilon_1 = (0,-1,3,-3) \sim 0.0988$

$a_5 := a_4 + \varepsilon_3 = (-5,4,-1,1) \sim 1.7422$

Hence, we stop with the approximation $a_4 = -\theta_1 + 3\theta_2 - 3\theta_3$.

The error of this approximation is $0.00121\ldots$.

# Galois Spectrum

Let $\zeta = \exp(2\pi i/16)$.

Then $\theta_1 = \zeta + \zeta^{-1}$, $\theta_2 = \zeta^2 + \zeta^{-2} = \sqrt{2}$, and $\theta_3 = \zeta^3 + \zeta^{-3}$.

$\mathbb{Q}(\theta_1)$ is a Galois extension of $\mathbb{Q}$.

Its Galois group is cyclic and is generated by $\tau\colon \zeta+\zeta^{-1} \mapsto \zeta^5 + \zeta^{-5}$.

## Galois-Spectrum: Fundamental Equation

For $a = \alpha_0 + \alpha_1\theta_1 + \alpha_2\theta_2 + \alpha_3\theta_3$ we have

$$\begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = \frac{1}{8} \begin{pmatrix} 2 & 2 & 2 & 2 \\ \theta_1 & -\theta_3 & -\theta_1 & \theta_3 \\ \theta_2 & -\theta_2 & \theta_2 & -\theta_2 \\ \theta_3 & \theta_1 & -\theta_3 & -\theta_1 \end{pmatrix} \begin{pmatrix} a \\ \tau(a) \\ \tau^2(a) \\ \tau^3(a) \end{pmatrix}.$$

$\rightarrow L_\infty(a) \leq \text{maxconj}(a)$

$\rightarrow a$ has signature $(+,-,-,+)$ if $\tau(a)$ is positive and

$$\tau(a) \geq \frac{\theta_1}{\theta_3}(|a| + |\tau^2(a)| + |\tau^3(a)|)$$

$\rightarrow$ similar assertions hold for other signatures.

# Cyclotomic Units

Elements of $E$: power products of small elements with signatures $(+,-,-,+),(+,-,+,-),(+,+,-,-)$.

We use cyclotomic units: let

$$\eta_j := \zeta^j \frac{1-\zeta}{1-\zeta^{2j+1}}, \quad j = 1,2,3.$$

$\eta_j$ is real and is a unit of $\mathbb{Z}[\zeta]$, i.e., the product of its Galois-conjugates is plus or minus one.

# Linear Programming

Find $k_1, k_2, k_3$ such that $\varepsilon = \prod_{j=1}^{3} \eta_j^{k_j}$ satisfies $|\tau(\varepsilon)| \geq 2\frac{\theta_1}{\theta_3}|\tau^j(\varepsilon)|$ for $j = 2, 3$, and $(1 + \frac{\theta_3}{\theta_1})|\tau(\varepsilon)| \leq 4M - 1$.

Then $L_\infty(\varepsilon) \leq M$ and $\varepsilon$ has signature $(+, -, -, +)$ or $(-, +, +, -)$, according to whether $\tau(\varepsilon)$ is positive or not.

Take logarithms:

$$\sum_{l=1}^{3} k_l(\log|\tau^j(\eta_l)| - \log|\tau(\eta_l)|) \leq -\log(2) - \log(\theta_1) + \log(\theta_3)$$

and

$$\sum_{l=1}^{3} k_l \log|\tau(\eta_l)| \leq \log(4M - 1) + \log(\theta_1) - \log(\theta_1 + \theta_3).$$

Minimize $\sum_{l=1}^{3} k_l \log|\eta_l|$ subject to these inequalities. (3 constraints and 3 variables.)

The resulting algorithm has worst case approximation error $O(1/M^3)$ which is optimal according to the lower bound mentioned before.

It runs in time $O(\log(M))$.

## General Algorithm

For $n \geq 5$ there is no signature technique.

However, one can replace the signature by a more complicated attribute based on the magnitude of the Galois-conjugates.

In an analogous way one can construct a set $E$ consisting of power products of cyclotomic units.

The corresponding exponents can be found by solving linear equations of size $2^{n-2}$.

The resulting algorithm has optimal worst case error $O(1/M^{2^{n-2}-1})$ for fixed $n$ .

# Experimental results

$n = 3$, $M \sim 54 \times 10^6$: 5000 complex approximations take 0.6 seconds on a SPARC-5.

$n = 4$, $M \sim 2 \times 10^5$: 13000 real approximations take 0.7 seconds on an ULTRASPARC-1.

$n = 5$, $M \sim 3 \times 10^3$: 22000 real approximations take 17 seconds on an ULTRASPARC-1.

In this case we machine-generated the approximation program and ran it without fine tuning.

For practical purposes the approximation algorithm for $n = 4$ gives a good tradeoff between accuracy and running time.

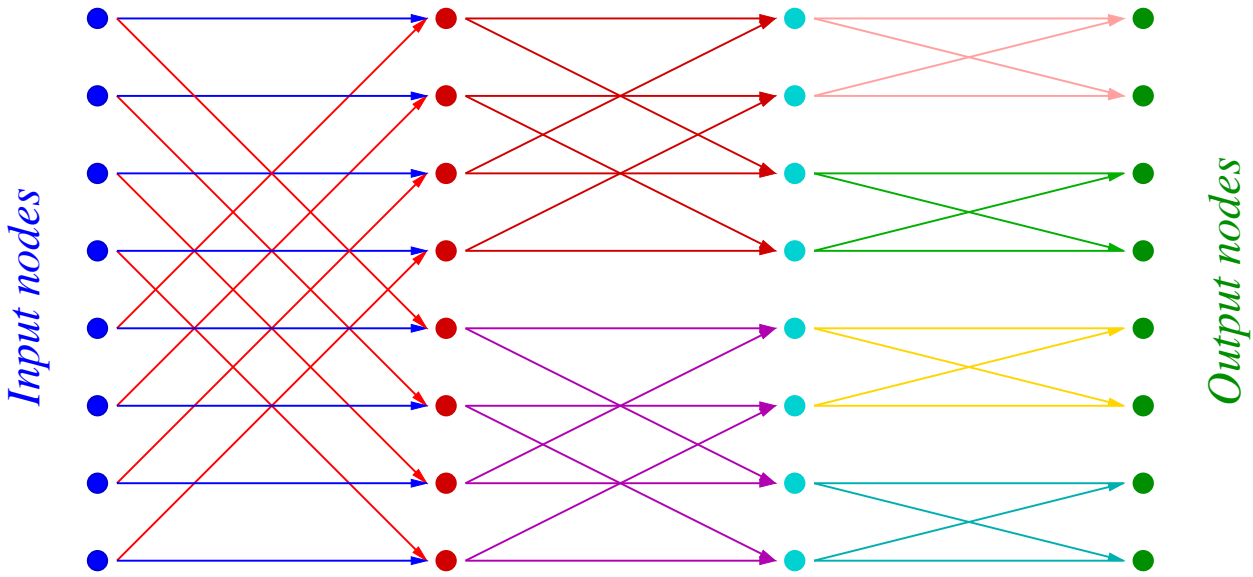
# Application: The Fast Fourier Transform

Let $N := 2^n$ and $\zeta := e^{2\pi i/N}$. The Discrete Fourier Transform of a complex vector $(a_0, \ldots, a_{N-1})$ is the vector $(\widehat{a}_0, \ldots, \widehat{a}_{N-1})$ defined by

$$\forall\, j = 0, \ldots, N-1: \quad \widehat{a}_j = \sum_{k=0}^{N-1} a_k \zeta^{jk}.$$

Straight-forward computation needs $O(N^2)$ arithmetic operations.

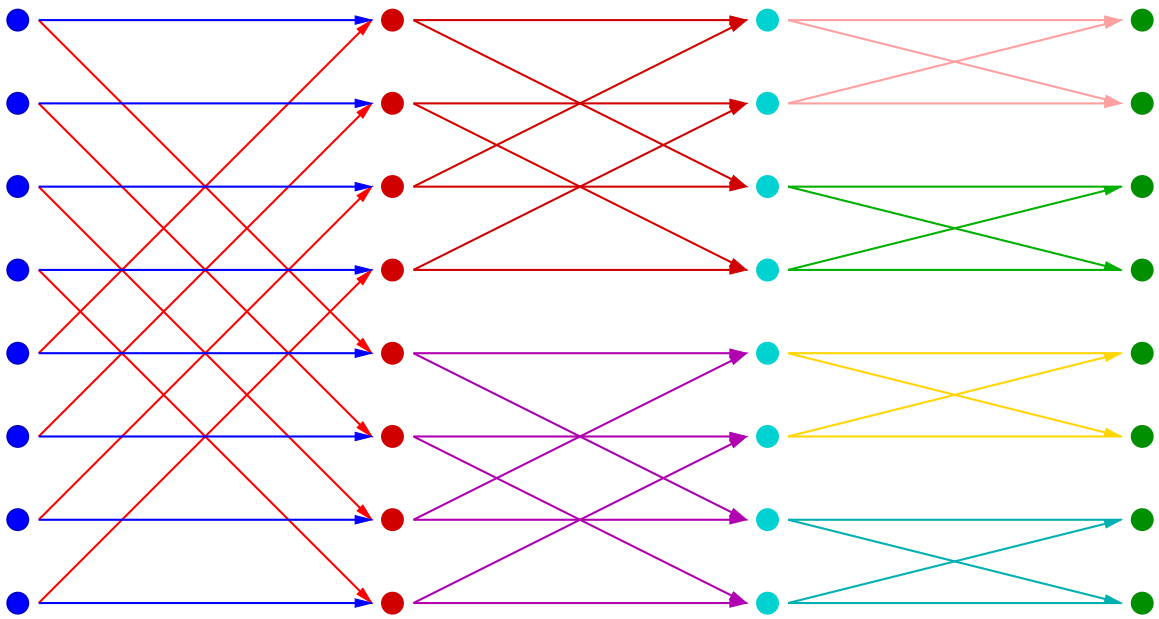
# Example: FFT of Length 8



# Precision

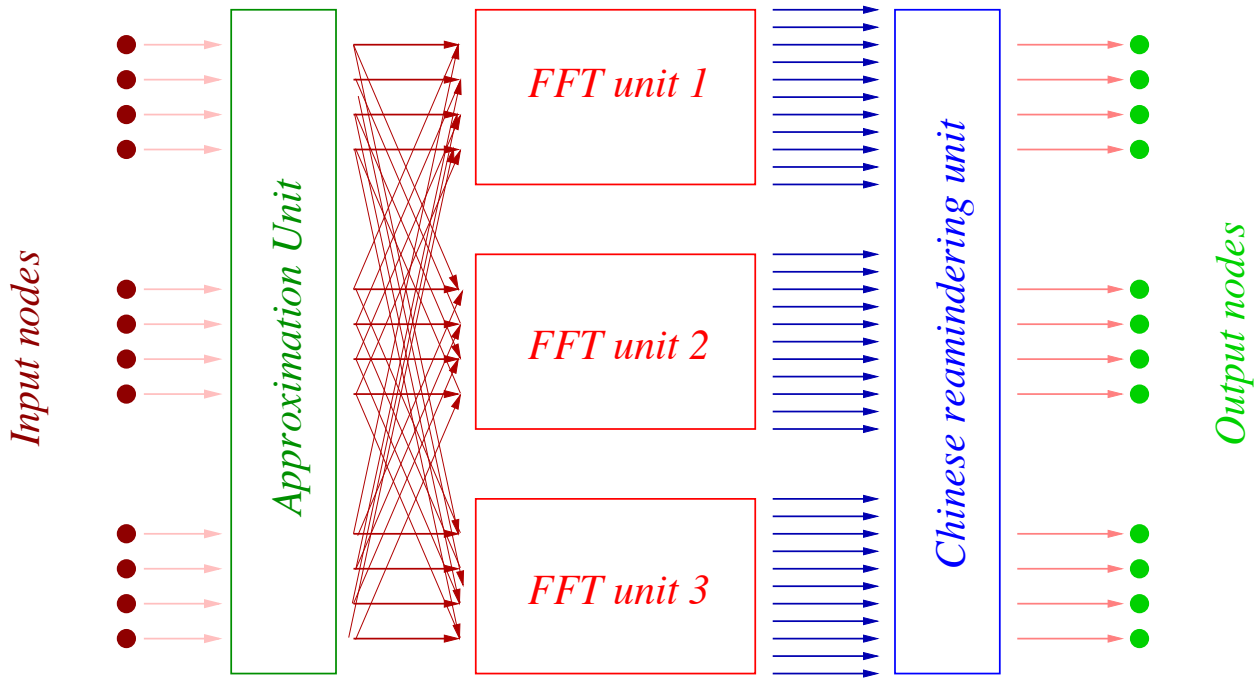
At each level of the FFT the inevitable scaling leads to the loss of 1/2 bits on average when working on a fixed point basis.

An FFT- accompanied by an inverse FFT of a vector of length 1024 leads to a loss of 10 bits. This can be fatal on a 16-bit fixed point processor.



# Fixed point FFT processor chart

The following algorithm was proposed by Cozzens and Finkelstein'85:




Main focus: Approximation unit.

# Chinese Remaindering

When multiplying Gaussian integers, the results become large. To avoid an overflow, one can use an ancient mathematical parallel processing technique known as Chinese remaindering.

For polynomials Chinese remaindering is nothing but the well-known evaluation and interpolation.

# Example

$$x + iy = (2 + 3i)(5 + 2i) \equiv 4 + 5i \bmod 7$$
$$\equiv 4 + 8i \bmod 11$$

$$x \equiv 4 \bmod 7 \qquad y \equiv 5 \bmod 7$$
$$\equiv 4 \bmod 11 \qquad \equiv 8 \bmod 11$$

---

$$x \equiv 4 \bmod 77 \qquad y \equiv 19 \bmod 77.$$

---

$$\underline{x = 4} \qquad \underline{y = 19}$$

## New Algorithm

We will present an algorithm that yields $b$-bit accurate FFT's in time

$$O(b \log(b) N \log(N)).$$

Note that the floating point algorithm yields $b$-bit accurate FFT's in time

$$O(b^2 N \log(N)).$$