Nanoscale Resonance Energy Transfer-Based Devices for Probabilistic Computing

Despite the theoretical advances in probabilistic computing, a fundamental mismatch persists between the deterministic hardware that traditional computers use and the stochastic nature of probabilistic algorithms. The authors propose Resonance Energy Transfer (RET) between chromophores as an enabling technology for probabilistic computing functional units. RET networks can implement efficient samplers with arbitrary probability distributions and have great potential for accelerating probabilistic algorithms.

Siyang Wang
Alvin R. Lebeck
Chris Dwyer
Duke University

Modern computers largely take a deterministic approach to computation and are designed with deterministic algorithms and transistor functionality in mind. However, recent challenges in CMOS scaling reveal practical limits on performance as lithographic features continue to shrink. Looking beyond the gains that technology scaling has traditionally supplied, recent theoretical advances in statistics and probabilistic machine learning have demonstrated that many application domains can benefit from probabilistic algorithms in terms of solution quality and efficiency.[1] However, a fundamental mismatch exists between the deterministic hardware that traditional computers use and the stochastic nature of these new algorithms. Furthermore, significant performance gains can be found by carefully employing alternative technologies embedded in conventional systems to better match the probabilistic algorithms' requirements.

Most probabilistic computations rely on sampling from an application-specific distribution of interest, which requires control over a parameterizable source of entropy used for random selection. Pseudo-random number generation—for example, by linear feedback shift registers—can be used in a deterministic framework to emulate a random process but incurs significant overhead per sampling operation (for example, hundreds of instructions). A promising alternative is to use a physical process that is a natural source of entropy.

The common physical processes used for this fall into three categories: thermal phenomena, the photoelectric effect, and quantum

Published by the IEEE Computer Society

0272-1732/15/$31.00 © 2015 IEEE

phenomena.[2] Researchers have proposed using thermal noise in electronic circuits to make probabilistic switches with a set of tunable parameters, but this approach requires amplifying the noise to a specific magnitude that can be energy and area inefficient.[3] Furthermore, probabilistic switches essentially implement Bernoulli random variables, which are not sufficiently general to scale and approximate arbitrary probabilistic behaviors. The photoelectric effect and quantum phenomena can implement many underlying distributions but are usually difficult to parameterize.[2]

Within this context, we present Resonance Energy Transfer (RET) as a new candidate technology for probabilistic computing, based on two enabling observations:

- Molecular implementations of RET networks with different geometries can produce output that is a direct, physical analog to different instances of phase-type distributions.
- Phase-type distributions can approximate general distributions and are easily adapted to any probabilistic computation.

This article focuses on using RET networks to implement new functional units dedicated to efficiently generating samples that can potentially accelerate probabilistic algorithms for different applications; a comprehensive architectural evaluation will be future work.

Phase-type distributions

To qualify as a building block for implementing functional units with arbitrary probabilistic behavior, the physical device's intrinsic probability distribution must be able to approximate general distributions. Examples of such probability distributions include mixtures of Gaussian distributions, Gamma distributions, and phase-type distributions,[4–6] each of which can form a dense set in the field of all positive-valued distributions and approximate any positive-valued distribution. The Poisson process and exponential distribution naturally exist in nanoscale processes (such as chemical reactions, intermolecular energy transfer, luminescence, and electrostatics) and, more importantly, can be flexibly convolved and mixed to form different instances of phase-type distributions. Thus, they provide unique opportunities for nanoscale physical implementation and approximating general distributions (see the sidebar "Continuous-Time Markov Chain and Phase-Type Distributions").

Molecular implementation using RET

RET networks can be designed to achieve different phase-type distributions. The RET transfer between two chromophores (single-molecule optical devices) is exponentially distributed in the time domain, and therefore a molecular scale RET network is a direct, physical analog of a phase-type distribution in which the RET network's geometry configures its corresponding phase-type distribution. RET networks can be conveniently and economically fabricated using DNA self-assembly with subnanometer precision.[7,8] Hence, RET technology becomes a natural substrate for implementing phase-type distributions.

RET

We propose a molecular implementation of phase-type distributions based on chromophores. A single chromophore absorbs photons of a specific wavelength and emits photons at a longer wavelength via fluorescence. However, when two chromophores are placed a few nanometers apart and their emission and excitation spectra overlap, energy transfer can occur between the two chromophores through RET. RET is a quantum mechanical energy-transfer mechanism between two chromophores, in which the donor chromophore, initially in its excited state, transfers its energy to the acceptor chromophore through nonradiative dipole-dipole coupling (see Figure 1).[9] The time to RET transfer, after the donor is excited, follows an exponential distribution between each chromophore pair; therefore, the sequence of RET transfers and the time from the exciton entering the chromophore network until leaving the network (that is, decaying) follows a phase-type distribution. The RET transfer between a chromophore pair with a transfer rate of k_{RET} physically implements a phase transition with a rate of $\lambda = k_{RET}$ in a phase-type distribution; the chromophore

Continuous-Time Markov Chain and Phase-Type Distributions

Within the context of Markov networks and Bayesian inference, the composition of several exponential distributions becomes an essential ingredient to make more interesting problem-specific joint distributions. However, many problems require distributions outside of this which do not have analytical closed forms and cannot be efficiently sampled. Phase-type distributions, which Resonance Energy Transfer (RET) networks physically realize, are a complete basis set from which any other distribution can be synthesized. Analogous to how a sine/cosine basis set can be used to synthesize any waveform by way of the Fourier transform, so can RET networks be used to physically transform a composition of phase-type distributions into any other distribution. Here, we describe the basic mathematical techniques used to analyze Markov networks and the series of states, or chains, they produce (often in time), and we briefly describe the well-understood connection to phase-type distributions.

Continuous-time Markov chain

A CTMC $X = \{X(t), t \geq 0\}$ is a continuous-time stochastic process with a finite or countable state space *S*, in which the time spent in each state is exponentially distributed. A CTMC's Markov property means that the conditional probability distribution of future states of the process (conditional on both past and present states) depends only on the present state and not on the sequence of events that preceded it. A CTMC is defined by its discrete state space *S*, a transition matrix *Q* that indicates the transition rate between each pair of states, and an initial probability distribution $\pi(0)$.

If a CTMC has at least one absorbing state (that is, a state from which there is no escape) $S_{Ai} (i = 1, \ldots, n)$, which only has incoming transition rates, the probability of the system being in an absorbing state approximates 1 as time increases to infinity: $\lim_{t\to\infty} Prob\left(X(t) \in \{S_{Ai}, i = 1, \ldots, n\}\right) = 1$. Additionally, the absorption probability of each absorbing state S_{Ai} $(i = 1, \ldots, n)$—that is, the probability of the system transitioning into this absorbing state—is $P_{Ai} = Prob(X(\infty) = S_{Ai})$, which is affected by the initial probability distribution $\pi(0)$ and the transition matrix *Q*.[1]

Phase-type distributions

Given an absorbing CTMC, the time to absorption *T* follows a phase-type distribution $f_T(t)$. The absorption of the system occurs after a sequence of states, the time spent in each state is exponentially distributed, and the sequence of traversed states before absorption is itself a random process. Because a phase-type distribution is a probability distribution constructed by a mixture and convolution of exponential distributions, we can represent it by the time to absorption of an absorbing CTMC that characterizes the exponentially distributed time in each phase and the transition between phases before absorption. For example, an *n*-stage Erlang distribution, a special instance of a phase-type distribution, can be perceived as the distribution of the time to absorption in the so-called pure birth process (that is, state transitions strictly move along one direction) with $n - 1$ transition states and one absorbing state at the end (see Figure A). Other special cases of phase-type distributions include exponential, Erlang, hyperexponential, and Coxian distributions.

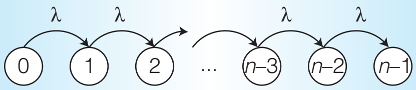

Figure A. A "pure birth" process with *n* states. This process is an example of a Markov [state] chain that follows in time (or discrete step) only one direction in the Markov network.

Approximating general distributions using phase-type distributions

Phase-type distributions are often used to approximate general distributions because a CTMC-based problem is often analytically tractable. In theory, any positive-valued discrete or continuous distribution can be approximated with a phase-type distribution to arbitrary precision. The two common approaches to approximating a general continuous distribution with a phase-type distribution are based on moment matching[2] and minimization of a difference (for example, Kullback–Leibler divergence).[3] However, the quality of the approximation in practice can be less than ideal for certain continuous distributions because of practical limits on the underlying phase-type distribution's type and size. Under such circumstances, discretizing the continuous distribution and then approximating the discretized distribution can improve the approximation.[4] Similar to the continuous case, a discrete phase-type distribution can be represented by the time to absorption of an absorbing discrete-time Markov chain and used to approximate a general discrete distribution.

References

1. K.S. Trivedi, *Probability & Statistics with Reliability, Queuing and Computer Science Applications*, John Wiley & Sons, 2008.
2. T. Osogami, and M. Harchol-Balter, *A Closed-Form Solution for Mapping General Distributions to Minimal PH Distributions*, Springer, 2003.
3. M. Olsson, *The Empht-Programme*, tech. report, Dept. Mathematics, Chalmers Univ. Technology, and Göteborg Univ., 1998.
4. A. Bobbio et al., "Acyclic Discrete Phase Type Distributions: Properties and a Parameter Estimation Algorithm," *Performance Evaluation*, vol. 54, no. 1, 2003, pp. 1–32.

network's geometry controls how these phases are convolved and mixed to form the phase-type distribution. Specifically, the RET transfer rate between each chromophore pair and the decay rate of each relaxation pathway constitute the transition matrix Q of the absorbing continuous-time Markov chain (CTMC) that defines the phase-type distribution.

The rate of the RET process between a given chromophore pair is

$$k_{RET} = \frac{3}{2}\frac{k^2}{\tau_D^0}\left(\frac{R_0}{r}\right)^6, \qquad (1)$$

where τ_D^0 is the donor's intrinsic fluorescence lifetime, k^2 is the mutual orientation of the chromophore pair, r is the free space distance between the chromophore pair, and R_0 is the Förster radius (that is, the distance at which the transfer efficiency is 50 percent). The Förster radius of a chromophore pair mainly depends on the properties of the two chromophores, such as the donor's quantum yield Φ_D^0 and emission spectra $I_D(\lambda)$ and the molar absorption coefficient of the acceptor $\varepsilon_A(\lambda)$.

A chromophore's intrinsic fluorescence lifetime τ^0 is determined by the rates of all the intrinsic relaxation pathways, including both the radiative pathway (that is, fluorescence) and nonradiative pathways. In the presence of RET to another chromophore, another relaxation pathway is added, which has the rate k_{RET} described in Equation 1. The relaxation event through each pathway is exponentially distributed with its associated decay rate, and, as a result, the chromophore's deexcitation is also exponentially distributed. Between a RET pair, the donor chromophore's excited state lifetime is shown in Equation 2, and the transfer efficiency is shown in Equation 3;

$$\tau_D = \frac{1}{1/\tau_D^0 + k_{RET}} \qquad (2)$$

$$TE = \frac{k_{RET}}{1/\tau_D^0 + k_{RET}} = \frac{1}{1+(r/R_0)^6}. \qquad (3)$$

The fundamental parameters that govern the transfer probability between chromophores are largely defined by the molecules we choose to create the network. However, the separation between chromophores can be precisely controlled through DNA self-assembly.[7,8] This process can create well-defined nanoscale networks of chromophores with hundreds of states and a widely tunable set of transfer probabilities.[10–12]

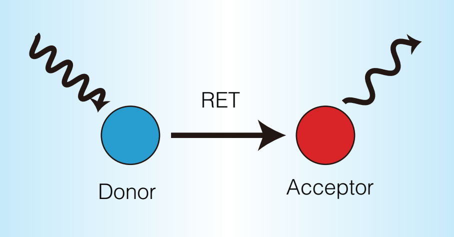

Figure 1. A diagram of Resonance Energy Transfer (RET). An incident photon excites the donor, which transfers its excited state energy to the acceptor via RET. The acceptor can then emit a longer wavelength photon.

A natural CTMC

Given that a chromophore network physically implements an absorbing CTMC, the time to exciton decay in the network follows a phase-type distribution. (Several assumptions are necessary here: that the pulsed light source excites only one donor chromophore in the system, at most one chromophore remains excited at any time, and nonlinear mixing is negligible. From our experience with real devices and networks, these are reasonable assumptions.) Figure 2 shows the energy transfer via RET and the excitation decay via fluorescence between a chromophore pair (Figure 2a) and its corresponding CTMC model (Figure 2b). In the CTMC, each chromophore has a transient state S_T to indicate whether it is in its excited state, and the exciton decay through each relaxation pathway of each chromophore is represented as an absorbing state S_A. Although it is not shown in Figure 2a, nonradiative decay exists as an exciton relaxation pathway and is included in Figure 2b. Equation 4a shows the initial state vector $\pi(0)$, and Equation 4b shows the transition matrix Q of the CTMC:

$$\pi(0) = [1, 0, 0, 0, 0, 0] \qquad (4a)$$

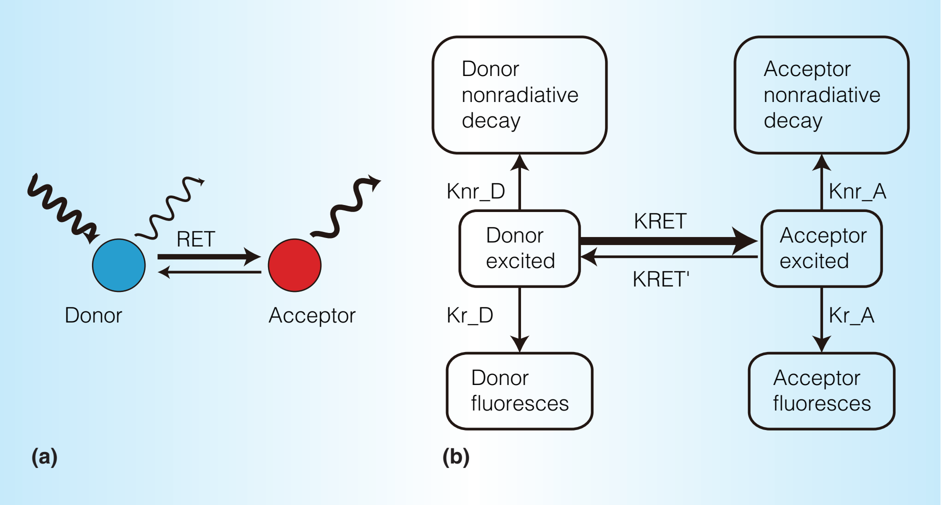

Figure 2. RET between a chromophore pair. (a) Energy transfer and decay processes. (b) The corresponding Markov network.

$$Q = \begin{array}{c} \\ \\ \\ \end{array}\left[\begin{array}{cccccc} S_{TD} & S_{TA} & S_{A1} & S_{A2} & S_{A3} & S_{A4} \\ -K_{RET}-K_{r\_D}-K_{nr\_D} & K_{RET} & K_{r\_D} & 0 & K_{nr\_D} & 0 \\ K_{RET'} & -K_{RET'}-K_{r\_A}-K_{nr\_A} & 0 & K_{r\_A} & 0 & K_{nr\_A} \\ \multicolumn{2}{c}{0_{4X2}} & \multicolumn{4}{c}{0_{4x4}} \end{array}\right]. \quad (4b)$$

Figure 3a shows the state probabilities of the four absorbing states in Equation 4b: $\pi_{SA}(t)=[\pi_{SA1}(t), \pi_{SA2}(t), \pi_{SA3}(t), \pi_{SA4}(t)]$ (S_{A1}: donor fluoresces, S_{A2}: acceptor fluoresces, S_{A3}: donor nonradiative decay, and S_{A4}: acceptor nonradiative decay). Their sum monotonically approximates 1, because the input exciton is increasingly likely to have decayed as time passes. Specifically, the gray in Figure 3a correspond to the two absorbing states for the fluorescence of the two chromophores (the donor molecule is named AF488 and the acceptor molecular is named AF594). By taking the derivative of the gray solid curves in Figure 3a and normalizing each, we can achieve the conditional probability density function (PDF) of the time to fluorescence from each chromophore, $f_T(t|X(\infty)=S_{A1})$ and $f_T(t|X(\infty)=S_{A2})$ (see Figure 3b). For a larger RET network with more than two chromophores, its CTMC will simply enclose more states and the transition rates between them (see Figure 4), and results similar to Figure 3 can be derived.

Using a RET circuit to sample from a general distribution

What remains to be shown is a systematic method by which probabilistic algorithms can exploit the natural random process in RET networks. We will describe the structure of a RET circuit and how to use it as a functional unit dedicated to generating samples for practical applications in probabilistic computing.

RET circuit

Once we design the network geometry and chromophore types, we can fabricate a RET network with subnanometer precision using hierarchical DNA assembly.[7,12,13] We can then deposit the fabricated network between an optical waveguide, which delivers an input pulse of light, and a photodetector, which detects the output from the network, to implement what we call a *RET circuit.*

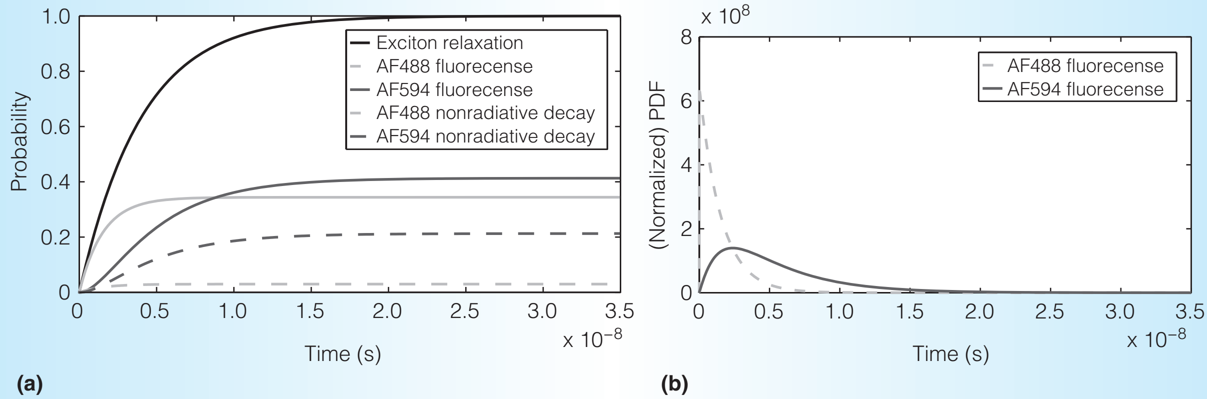

Figure 3. The continuous-time Markov chain (CTMC) solution for a chromophore pair. (a) The state probabilities of the four absorbing states in the Markov network in Figure 2b and their sum. (b) The conditional probability density function (PDF) of the time-resolved fluorescence from each chromophore. This theoretical derivation of the time-dependent evolution of the CTMC is a precise description of the observable physical process of RET.

Figure 5 illustrates a proposed RET circuit fabricated on top of a CMOS substrate that uses quantum dot LEDs (QD-LEDs) as a light source coupled through a waveguide to a RET network with multiple donors and a single accepter, and a single photon avalanche detector (SPAD) to detect fluorescence. This RET circuit shows a RET network with two RET pairs, two donors, and one acceptor (more pairs per network are possible). A RET network is very small and can reside in a thin layer (less than 20 nm × 20 nm × 2 nm) above the SPAD. The QD-LEDs also require a very small area $(0.15\,\mu\text{m}^2)$;[14,15] thus, the SPAD (about $100\,\mu\text{m}^2$) dominates RET circuit area.[16,17] RET circuits can be used to implement probabilistic functional units integrated with conventional CMOS.

Continuous distribution

Chromophore fluorescence occurs only when the network finds an absorbing state in the RET-based CTMC. The time to fluorescence (TTF) after an initial input pulse follows a phase-type distribution, which we can use to implement a general continuous distribution. The SPAD provides photon detection events, and the histogram of these event times approximates the PDF of the TTF. Therefore, a RET circuit implements a functional unit that can generate samples from a general continuous distribution.

With the RET pair in Figure 2, the PDF of the TTF at the wavelength of the donor fluorescence is the dotted curve and the PDF of the TTF at the wavelength of the acceptor fluorescence is the solid curve in Figure 3b. When the back transfer from the acceptor to the donor is negligible, the PDF of the donor TTF is an exponential distribution and the PDF of the acceptor TTF is a two-phase hypoexponential distribution. A functional unit with this RET circuit would sample from the two-phase distribution.

Discrete distributions

In addition to continuous distributions, discrete distributions are often necessary in common probabilistic algorithms. For example, a continuous distribution can be made discrete and then approximated to improve accuracy when the phase-type distribution is constrained in size. Similar to its continuous counterpart, a discrete phase-type distribution can be defined by a discrete-time Markov chain (DTMC) and can approximate a general discrete distribution. A DTMC can be constructed conveniently by discretizing the continuous time domain of a CTMC with a time increment Δt. Hence, the RET-based CTMC can be used to construct a DTMC that approximates a general discrete distribution.

We explored two approaches to implement a discrete distribution with RET-based

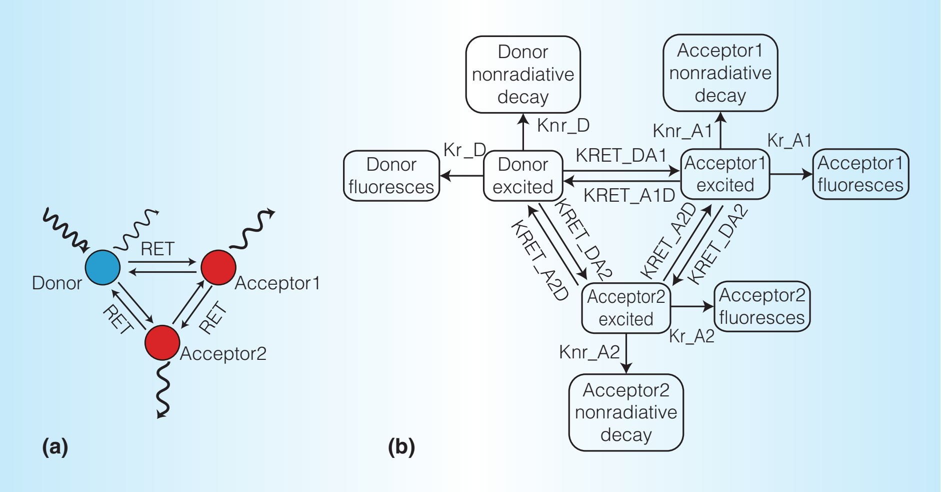

Figure 4. A larger RET network. (a) Energy transfer and decay processes. (b) The corresponding Markov network.

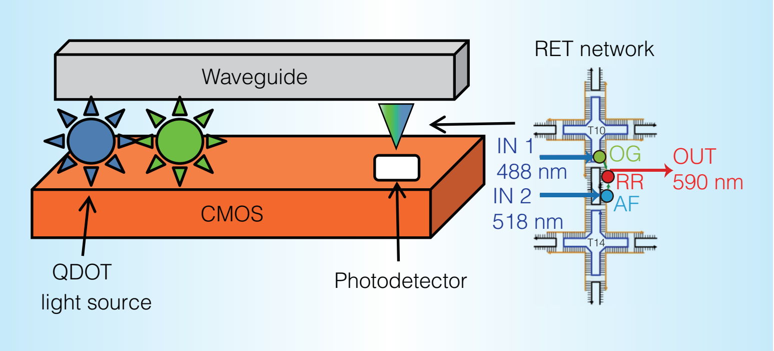

Figure 5. Illustration of a proposed RET circuit fabricated on top of a CMOS substrate. CMOS-integrated photonics are an important enabling aspect of this work.

phase-type distributions for Bernoulli and exponential functional units. They both rely on a small-scale RET network, are convenient to design, and could serve as the basis for a discrete sampler useful in many probabilistic algorithms.

Implementation 1: Bernoulli functional units. We could implement a Bernoulli random variable using a RET circuit by observing the presence or absence of photon detection events within a specified time period after excitation. Consider a single chromophore pair excited by one QD-LED at time $t = 0$. Given a detection interval $[0, T]$, the SPAD outputs "1" if it detects any photons within this interval, and "0" otherwise. When the back transfer from the acceptor to the donor is negligible, the probability that the SPAD fails to detect any photons from the donor in the interval is

$$P_B = [1 - P_e P_{DF} P_d (1 - e^{-\lambda_D T})]^N \tag{5a}$$

$$\lambda_D = \lambda_{d0} + k_{RET} = k_{d_r} + k_{d_{nr}} + k_{RET} \tag{5b}$$

$$P_{DF} = \frac{k_{d_r}}{k_{d_r} + k_{d_{nr}} + k_{RET}}. \tag{5c}$$

In Equation 5, λ_D is the decay rate of the excited state of the donor chromophore in the presence of RET to the acceptor, which is the sum of its intrinsic excited state decay rate λ_{d0} and the RET transfer rate k_{RET}. P_{DF} is the probability that the donor fluoresces after it is excited, which is the ratio between its radiative decay rate k_{d_r} and $\lambda_D = k_{d_r} + k_{d_{nr}} + k_{RET}$. P_e is the probability of the donor being excited by the QD-LED, and P_d is the SPAD's photon detection efficiency. The cumulative distribution function of an exponential distribution in the inner parenthesis accounts for the donor's TTF, and taking the complement of the probability $P_e P_{DF} P_d (1 - e^{-\lambda_D T})$ results from the fact that we are evaluating the probability of not detecting photons from a single RET network in the time interval $[0, T]$. We assume that many copies of the RET network are integrated in a RET circuit for ease of fabrication and reliability, and N is the number of RET networks in the ensemble that can be excited. This is sufficient to implement a

Bernoulli random variable with the parameter $p = P_B$.

Given a certain configuration of the other parameters, Figure 6 shows the dependence of P_B of a RET circuit on the RET transfer rate k_{RET}; P_B is approximately 0 when $k_{RET} = 0$ and monotonically approximates 1 when k_{RET} increases. As a result, the RET transfer rate k_{RET} alone is enough for reaching different values of p in the range (0,1), even when the other parameters are fixed or limited (for example, fixed QD-LED intensity, limited types of available chromophores, fixed number of RET networks in the ensemble, and fixed time interval T).

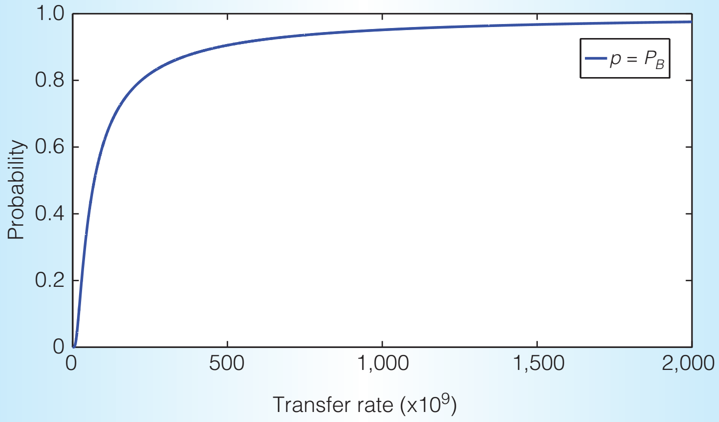

Figure 6. The parameter p of an implemented Bernoulli random variable depends on the RET transfer rate k_{RET}. The ability to precisely control k_{RET} implies that many p can be realized.

Therefore, we can use a RET circuit as a Bernoulli functional unit that generates binary samples with a parameter p. Multiple Bernoulli functional units can be composed to create higher-level functional units for discrete distributions in the same way that probabilistic switches are used to implement discrete distributions.[3] Consider a discrete random variable X with M possible outcomes $\{0, 1, \ldots, M - 1\}$ with the probabilities $\{P(X = 0) = p_0, P(X = 1) = p_1, \ldots, P(X = M - 1) = p_{M-1}\}$ $(p_0 + p_1 + \cdots + p_{M-1} = 1)$. The discrete random variable X can be implemented using M Bernoulli random variables $X_0, X_1, \ldots, X_{M-1}$ that correspond to the M outcomes $0, 1, \ldots, M{-}1$. The parameters of the M Bernoulli random variables $X_0, X_1, X_2, \ldots, X_{M-1}$ are, respectively, $p_0, \frac{p_1}{1-p_0}, \frac{p_2}{1-p_0-p_1}, \ldots, \frac{p_{M-1}}{1-\sum_{i=0}^{M-2} p_i}$. When $X_0, X_1, \ldots, X_{M-1}$ are sequentially sampled, the outcome corresponding to the first Bernoulli random variable whose sample is "1" is considered one sample of X.

Implementation 2: exponential functional units. A RET network's TTF is a phase-type distribution, and the SPAD can be used to sample from the phase-type distribution. The exponential distribution is simply a one-stage phase-type distribution; thus, it can conveniently be extracted from a RET network. Consider a single RET pair excited by one QD-LED at time $t = 0$. When the back transfer from the acceptor to the donor is negligible, the donor's TTF is exponentially distributed with a decay rate λ_D (Equation 5b). Because the SPAD detects photons beginning at $t = 0$, the time to the first photon detection has an exponential distribution with a decay rate λ:

$$\lambda = NP_eP_{DF}P_d\lambda_D = NP_eP_dk_{d_r} \quad (6a)$$

$$\lambda_D = \lambda_{d0} + k_{RET} = k_{d_r} + k_{d_{nr}} + k_{RET} \quad (6b)$$

$$P_{DF} = \frac{k_{d_r}}{k_{d_r} + k_{d_{nr}} + k_{RET}}, \quad (6c)$$

where all the parameters are the same as in Equation 5. Although λ is independent of the RET transfer rate, its value can still be engineered by changing the concentration of RET networks (N), the emission intensity of the QD-LED (P_e), and even the donor chromophore (k_{d_r}). In addition, as we will explain, we can use a set of such exponential distributions to implement a target discrete distribution, and only the relative ratio between their decay rates, rather than their exact values, matters.

We can use a RET circuit as an exponential functional unit that generates exponentially distributed samples. Multiple exponential functional units can be composed to create higher-level functional units for discrete distributions based on the property of competing exponential random variables.[18] Given M exponential random variables X_i $(i = 1, \ldots, M)$ with decay rates $\lambda_i (i = 1, \ldots, M)$, the probability the ith exponential random variable is the minimum among all the M random variables is

$$P\Big(X_i = \min(X_1, \ldots, X_M)\Big) = \frac{\lambda_i}{\sum_{j=1}^{M} \lambda_j}. (7)$$

Consider a discrete random variable X with M possible outcomes $\{0, 1, \ldots, M-1\}$ with the probabilities $\{P(X=0)=p_0, P(X=1)=p_1, \ldots, P(X=M-1)=p_{M-1}\}$ $(p_0+p_1+\ldots+p_{M-1}=1)$. The discrete random variable X can be implemented using M RET-based exponential random variables $X_0, X_1, \ldots, X_{M-1}$ that correspond to the M outcomes 0, 1, …, $M-1$. The decay rates of the M exponential random variables $X_0, X_1, \ldots, X_{M-1}$ are such that their relative ratio equals the relative ratio between $p_0, p_1, \ldots, p_{M-1}$ (that is, $\lambda_0 : \lambda_1 : \ldots : \lambda_2 = p_0 : p_1 : \ldots : p_{M-1}$). We can generate samples of the M-outcome discrete random variable X as follows. At time $t=0$, the QD-LEDs in the M RET circuit simultaneously send a delta pulse to excite their RET networks, and all the SPADs are turned on simultaneously. We consider the outcome corresponding to the RET circuit whose SPAD detects the first photon before all the other RET circuits to be a sample of X. The functional unit input is the signal to turn on the QD-LEDs and start the SPADs, whereas the output is the sample of X.

Suitable probabilistic algorithms

Repeated random sampling from known distributions is the core of Monte Carlo methods, which are widely used for optimization, numerical integration, and sampling from a target probability distribution. Specifically, Monte Carlo Markov chain (MCMC), a subset of Monte Carlo methods, is a critical approach to solving Bayesian inference problems. When very many samples are required, directly generating samples from RET-based functional units can accelerate Monte Carlo methods in the absence of the overhead required to computationally generate samples from a continuous or discrete distribution (for example, inverse transform sampling and rejection sampling). In addition, designing hardware based on generating samples for probabilistic algorithms could bring savings in energy efficiency and circuit area.[3,18] Moreover, the natural entropy used to generate random samples from a RET circuit makes it a natural, or *true,* random number generator that could suffer less from distributional bias.

Many probabilistic algorithms can potentially benefit from RET circuits, including Bayesian networks, probabilistic cellular automata, and random neural networks. Although these probabilistic algorithms can use the samples generated in RET circuits, the required sampling distributions and the procedure for exploiting the generated samples to achieve the final result often depend on the target probabilistic algorithm. Here, we show two examples of using RET circuits for probabilistic algorithms—*f*-divergence-based model selection and inference in a Bayesian network—to illustrate two different approaches to incorporating RET-based sample-generating functional units into probabilistic algorithms.

Case study: *f*-divergence-based model selection

Given two model distributions $f_{M1}(x)$ and $f_{M2}(x)$ and an input set of observations generated from an unknown true distribution $x_i \sim f_M(x)(i = 1, \ldots, N)$, a common task in statistics is to select the model that best describes the observations. We can calculate the goodness of fit in several ways depending on our interpretation of what distance measures are sensible for the problem. Likelihood measures are common and calculated by $\prod_{i=1}^{N} f_{M1}(x_i)$ and $\prod_{i=1}^{N} f_{M2}(x_i)$ for all observations and models. The model with a higher likelihood is most likely the model from which the observations were generated. We can show that, in the limit of an infinite number of observations, the likelihood ratio test—that is, $\prod_{i=1}^{N} f_{M1}(x_i)/\prod_{i=1}^{N} f_{M2}(x_i)$—is equivalent to calculating the difference in the Kullback–Leibler (KL) divergence, a measure of entropic dissimilarity, between the true distribution M and each model distribution $D_{KL}(M|\,M1) - D_{KL}(M|M2)$. As a result, the likelihood-based model selection essentially locates the model that has the minimum KL divergence from the true distribution.

Unfortunately, the likelihood calculation can be computationally expensive, particularly as the number of observations increases.

```
S_M1=0; S_M2=0;                          // initial scores of M1 and M2 start at 0
for(i=1:N) {                             // iterate over all observations
        do {
           t_M1 ~ f_M1(t); t_M2 ~ f_M2(t);    // generate one sample from each model RET circuit
           if(t_M2 == t_i)
               S_M1++;                   // M1 matches the observation
           if(t_M2 == t_i)
               S_M2++;                   // M2 matches the observation
        } while (t_M1 != t_i) && (t_M2 != t_i);   // continue until at least one model matches
} // the model with the largest score is most likely closer to the real distribution
```

Figure 7. One possible implementation of a model selection algorithm that exploits sampling from a RET circuit.

Within this context, we present an alternative approach to model selection that uses RET-based functional units to generate samples and needs only minimal numerical computation. In practical applications, sensors can collect observations and map them to the time domain through specific techniques such as amplitude-to-time conversion.[19] Two RET circuits implement two continuous model distributions, $f_{M1}(t)$ and $f_{M2}(t)$. With N observations $t_i (i = 1, \dots N)$, we can locate the model closer to the true distribution $f_M(t)$ behind these observations by repeatedly sampling from the two model RET circuits as shown in Figure 7.

The scores of the two models are zero at the beginning. With each input observation t_i, each of the two model RET circuits is excited by a pulse to generate a sample: t_{M1} and t_{M2}. Because of the limited precision of photon detection time, the samples are discretized and fall into narrow bins in the time axis. When at least one of the two samples t_{M1} and t_{M2} matches the current observation t_i, the score of the corresponding model is incremented by 1, and it moves on to the next input observation t_{i+1}. Otherwise, two new samples t_{M1} and t_{M2} are generated until at least one of them matches the current observation t_i. After all the N input observations are evaluated, the model with the higher score is considered closest to the true distribution behind these observations.

The mathematical proof for this sampling-based approach is straightforward (see the sidebar "A New f-Divergence Measured through Sampling"). Because KL divergence is also an f-divergence, with the f function being $f(t) = t \ln t$, this method simply uses another f-divergence for model comparison. But unlike the likelihood-based method, the method we present here is mainly based on sampling and requires minimal numerical computation.

Case study: inference in a Bayesian network

A Bayesian network is a probabilistic graphical model that describes the conditional dependencies between a set of random variables via a directed acyclic graph (DAG). It has important applications in various fields, including medical diagnosis, computer vision, natural language processing, and weather forecasting. Given a Bayesian network describing a problem, the joint probability distribution of all random variables can be derived, and unobserved random variables can be inferred. For example, the classic "sprinkler problem"[1] is a Bayesian network that models the belief that on any given cloudy day (or not), either rain or a sprinkler is the cause of any observed wet grass. The task at hand is to then calculate the joint distribution and conditional distributions over the various binary states of the system: C (cloudy), S (sprinkler), R (rain), and W (wet grass). Each of the various conditional distributions—for example, that the wet grass is due to only the sprinkler, $P(S = \text{true} \mid W = \text{true}, R = \text{false})$—is calculated using the conditional probability tables (CPTs) for each random variable. The CPTs are

A New *f*-Divergence Measured through Sampling

The input observations follow the true distribution: $t_i \sim f_M(t)$ $(i = 1, \ldots, N)$. With each of these observations t_i, two samples t_{M1} and t_{M2} are generated from the two model Resonance Energy Transfer circuits. The probability that $t_{M1}\,(t_{M2})$ is in the same time bin with t_i is $f_{M1}(t_i) * \Delta t (f_{M2}(t_i) * \Delta t)$, where Δt is the width of each time bin. When the time bin is narrow enough, the probability that t_{M1} and t_{M2} both match t_i is negligible. As a result, the probability that $M1$ receives one increment is $f_{M1}(t_i)/(f_{M1}(t_i) + f_{M2}(t_i))$, and the probability that $M2$ receives one increment is $f_{M2}(t_i)/(f_{M1}(t_i) + f_{M2}(t_i))$. When the number of input observations is large enough, the difference between the scores of the two models is shown in Equation A:

$$S_{M1} - S_{M2} \approx N * \int f_M(t) \frac{f_{M1}(t) - f_{M2}(t)}{f_{M1}(t) + f_{M2}(t)} dt \qquad (A)$$

$$\begin{aligned} S_{M1} - S_{M2} &\approx N * \int f_M(t) \frac{f_{M1}(t) - f_{M2}(t)}{f_{M1}(t) + f_{M2}(t)} dt \\ &\approx N * \int f_{M1}(t) * \frac{\dfrac{f_{M1}(t)}{f_{M2}(t)} - 1}{\dfrac{f_{M1}(t)}{f_{M2}(t)} + 1} dt \qquad (B) \\ &\approx N * \int \frac{f_{M1}(t)}{f_{M2}(t)} * \frac{\dfrac{f_{M1}(t)}{f_{M2}(t)} - 1}{\dfrac{f_{M1}(t)}{f_{M2}(t)} + 1} * f_{M2}(t) dt \end{aligned}$$

When $M1$ is the true distribution $f_M(t) = f_{M1}(t)$, Equation B shows that $(S_{M1} - S_{M2})/N$ approximates the *f*-divergence between $f_{M1}(t)$ and $f_{M2}(t)$, where the *f* function is $f(t) = t((t-1)/(t+1))$. From the properties of *f*-divergence, $(S_{M1} - S_{M2})$ is nonnegative and equals 0 only when $f_{M1}(t)$ and $f_{M2}(t)$ are identical. Similarly, when $M2$ is the true distribution $f_M(t) = f_{M2}(t)$, then $(S_{M2} - S_{M1})$ is nonnegative and equals 0 only when the models are identical. When the true distribution is between the two models, the scores of the two models reflect the divergence between the true distribution and each model distribution. Therefore, this model-selection method can select the better model given enough observations, and it is based on the new *f*-divergence, which can be measured through repeated sampling.

individually derived from various beliefs, or even data, about the underlying process (that is, the weather and lawn maintenance). For example, the CPT for the R state relates the probability $P(R = \text{true})$ and $P(R = \text{false})$ as a function of $C = \text{true}$ and $C = \text{false}$.

Instead of numerically calculating the joint distribution $P(C, S, W, R)$ and conditional distributions for different queries, such as $P(C)$, $P(S|W)$, and $P(S|W, R)$, using variable elimination and belief propagation, we can generate samples from the Bayesian network to fulfill these tasks.[18] We can generate a sample $[c, s, w, r]_i$ from the joint distribution by sequentially sampling each random variable given its parents using the CPT associated with this random variable in the order of $C \rightarrow \{S, R\} \rightarrow W$. To implement a discrete sampler functional unit, a row of RET circuits can be fabricated to match each row in the CPT of a random variable and generate a sample for the random variable given its parents' values. With a large number of samples $[c, s, w, r]_i$ $(i = 1, \ldots, N)$ iteratively generated this way, the joint and conditional distributions can be approximated by analyzing these samples. For example, to approximate $P(S|W = T, R = F)$, all the m samples $[c, s, w, r]_{k_i} (1 \leq k_1 < \cdots < k_m \leq N)$ that satisfy $w = T$ and $r = F$ are extracted, and within the m samples the percentage of samples with $s = T$ (or $s = F$) approximates $P(S = T|W = T, R = F)$ or $P(S = F|W = T, R = F)$.

Although the sprinkler problem is simple, large-scale and complex Bayesian networks are used in real-world applications where the numerical approach to inferring random variables becomes more computationally expensive. For such applications, the sampling approach combined with RET-based functional units dedicated to efficiently generating samples becomes an attractive new direction.

The implication of precise molecular-scale self-assembly and the control of physically stochastic quantum processes on computing has only recently begun to be studied. Unlike classical quantum computing or communication systems that rely on delicate quantum states, RET circuits have demonstrated robust, long-lived molecular states and offer an opportunity to exploit quantum effects (other than entanglement) in a practical fashion for probabilistic computing. Many challenges remain, specifically in demonstrating the integration of molecular RET circuits with existing on-chip photonic technology. Important measures of feasibility that are often ignored in basic science—such as the RET circuit longevity, microarchitecture, and sampler-to-memory bandwidth—must all be ironed out before any conclusive statement about relative performance can be made.

We have presented RET as a promising candidate technology for probabilistic computing and shown the early evidence that it may be a good hardware match to such problems. A full architectural evaluation of performance, power, and area is planned as future work. MICRO

References

1. K.P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, 2012.
2. R.L. Liboff, *Introductory Quantum Mechanics*, 4th ed., Addison-Wesley, 2003.
3. P. Korkmaz, "*Probabilistic CMOS (PCMOS) in the Nanoelectronics Regime*," PhD dissertation, School of Electrical and Computer Eng., Georgia Institute of Technology, 2007.
4. K.N. Plataniotis and D. Hatzinakos, "Gaussian Mixtures and Their Applications to Signal Processing," *Advanced Signal Processing Handbook*, S. Stergiopoulos, ed., CRC Press, 2000, pp. 3-1–3-35.
5. A. Mohammadi, M. Salehi-Rad, and E. Wit, "Using Mixture of Gamma Distributions for Bayesian Analysis in an M/G/1 Queue with Optional Second Service," *Computational Statistics*, vol. 28, no. 2, 2013, pp. 683–700.
6. D.R. Cox, "A Use of Complex Probabilities in the Theory of Stochastic Processes," *Mathematical Proc. Cambridge Philosophical Soc.*, vol. 51, no. 2, 1955, pp. 313–319.
7. C. Pistol and C. Dwyer, "Scalable, Low-Cost, Hierarchical Assembly of Programmable DNA Nanostructures," *Nanotechnology*, vol. 18, no. 12, 2007, pp. 125,305–125,309.
8. C. LaBoda, H. Duschl, and C.L. Dwyer, "DNA-Enabled Integrated Molecular Systems for Computation and Sensing," *Accounts of Chemical Research*, vol. 47, no. 6, 2014, pp. 1816–1824.
9. B. Valeur, *Molecular Fluorescence: Principles and Applications*, Wiley, 2002.
10. C. Pistol et al., "Architectural Implications of Nanoscale-Integrated Sensing and Computing," *IEEE Micro*, vol. 30, no. 1, 2010, pp. 110–120.
11. M.D. Mottaghi and C. Dwyer, "Thousand-Fold Increase in Optical Storage Density by Polychromatic Address Multiplexing on Self-Assembled DNA Nanostructures," *Advanced Materials*, vol. 25, no. 26, 2013, pp. 3593–3598.
12. C. Pistol et al., "Encoded Multichromophore Response for Simultaneous Label-Free Detection," *Small*, vol. 6, no. 7, 2010, pp. 843–850.
13. C. Pistol, C. Dwyer, and A.R. Lebeck, "Nanoscale Optical Computing using Resonance Energy Transfer Logic," *IEEE Micro*, vol. 28, no. 6, 2008, pp. 7–18.
14. T. Kümmell et al., "Electrically Driven Room Temperature Operation of a Single Quantum Dot Emitter," *Proc. SPIE*, 2009; doi:10.1117/12.808165.
15. F. Hargart et al., "Electrically Driven Quantum Dot Single-Photon Source at 2 GHz Excitation Repetition Rate with Ultra-Low Emission Time Jitter," *Applied Physics Letters*, vol. 102, no. 1, 2013, pp. 011,126–011,126.
16. A. Rochas, P.-A. Besse, and R.S. Popovic, "Actively Recharged Single Photon Counting Avalanche Photodiode Integrated in an Industrial CMOS Process," *Sensors and*

Actuators A: Physical, vol. 110, no. 1, 2004, pp. 124–129.

17. D. Palubiak et al., "High-Speed, Single-Photon Avalanche-Photodiode Imager for Biomedical Applications," *IEEE Sensors J.*, vol. 11, no. 10, 2011, pp. 2401–2412.
18. V.K. Mansinghka and E. Jonas, "Building Fast Bayesian Computing Machines Out of Intentionally Stochastic, Digital Parts," *Computing Research Repository* (ArXiv), 2014; arXiv:1402.4914.
19. B. Jong Cheol et al., "An Amplitude-to-Time Conversion Technique Suitable for Multichannel Data Acquisition and Bioimpedance Imaging," *IEEE Trans. Biomedical Circuits and Systems*, vol. 7, no. 3, 2013, pp. 349–354.

Siyang Wang is a PhD candidate in the Department of Electrical and Computer Engineering at Duke University. His research interests include molecular-scale computing and probabilistic computing. Wang has an MS in electrical and computer engineering from Duke University. He is a student member of IEEE. Contact him at siyang.wang@duke.edu.

Alvin R. Lebeck is a professor in the Departments of Computer Science and Electrical and Computer Engineering at Duke University. His interests include architectures for emerging nanotechnologies, multicore processors, memory systems, and energy-efficient computing. Lebeck has a PhD in computer science from the University of Wisconsin–Madison. He is a senior member of IEEE and a member of the ACM. Contact him at alvy@cs.duke.edu.

Chris Dwyer is an associate professor in the Departments of Computer Science and Electrical and Computer Engineering at Duke University. His research interests include DNA self-assembly and applications that expand the computational domain with a focus on device-to-systems design, evaluation, and synthesis. Dwyer has a PhD in computer science from the University of North Carolina at Chapel Hill. He is a senior member of IEEE and a member of the ACM. Contact him at c.dwyer@duke.edu.