

ТРАССИРОВКА ЛУЧЕЙ НА CUDA

Докладчик:

Фролов В.А. (ВМиК МГУ, G&M Lab; Nvidia)

Научный руководитель:

<u>Игнатенко А.В. (ВМиК МГУ, G&M Lab)</u>

Лекторы:

Боресков А.В. (ВМиК МГУ)

Харламов А.А. (ВМиК МГУ, Nvidia)



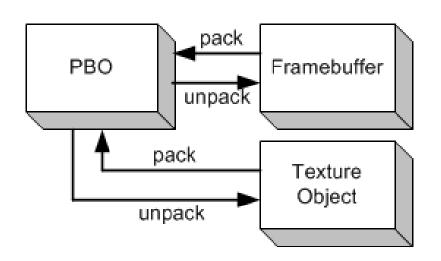
План

- ** OpenGL interoperability
- **Ж** Пересечение луча и треугольника
- **ж** Ускоряющие структуры
- **Ж** Организация трассировки лучей
- # Ray marching
- **ж** Задание



OpenGL interop

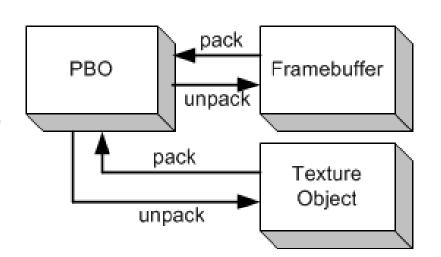
- # OpenGL 2.0
 - ✓ VBO, PBO
- # OpenGL 3.0
 - ✓ VBO, PBO
 - texture
 - renderbuffer
- # OpenGL 4.0
 - <u>∧</u> ?





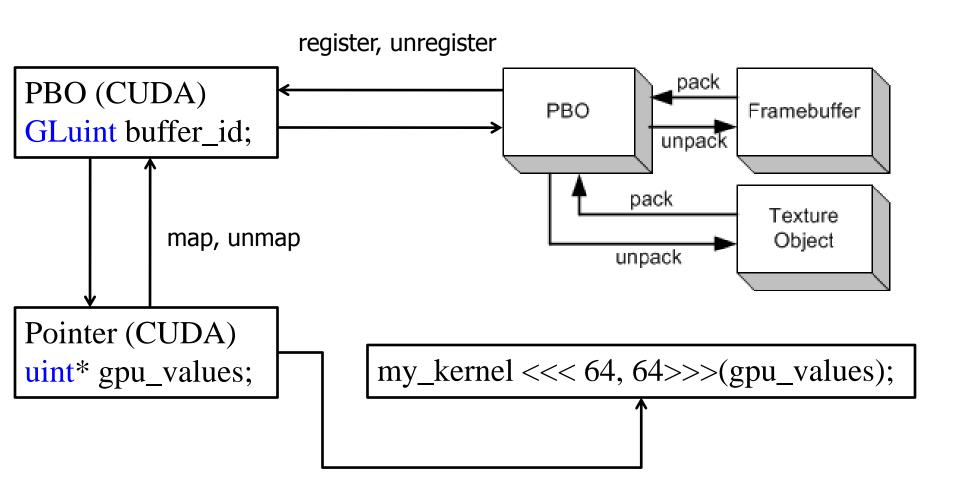
OpenGL 2.0 interop

- # Pack
 - □ glReadPixels (...)
 - glGetTexImage (...)
- # Unpack
 - □ glDrawPixels (...)
 - □ glTexImage2D (...)
 - glTexSubImage2D (...)





OpenGL 2.0 interop



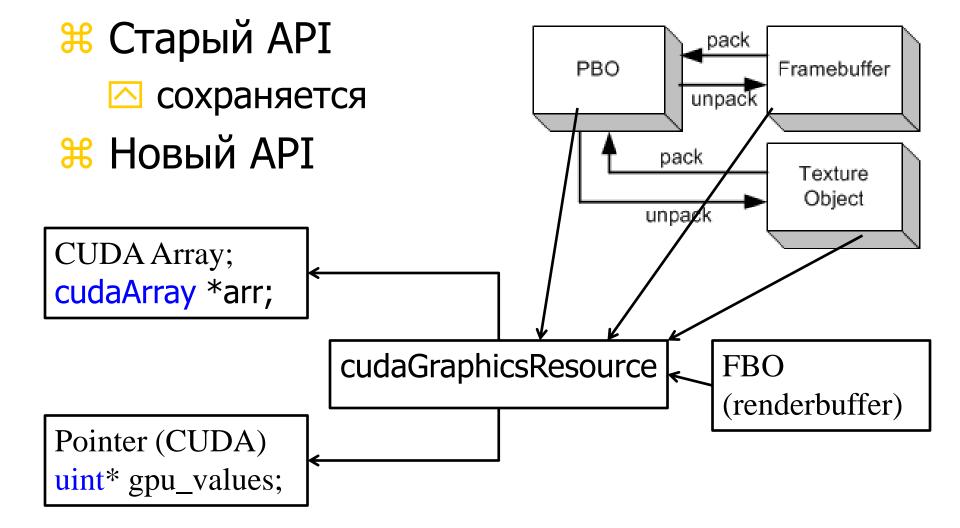


OpenGL 2.0 interop

```
glGenBuffers(1, &pixelBuffer);
glBindBuffer (GL_PIXEL_UNPACK_BUFFER, pixelBuffer);
glBufferData(GL_PIXEL_UNPACK_BUFFER, w*h*sizeof(int),
               screen_buffer, GL_STATIC_DRAW);
cudaGLRegisterBufferObject (pixelBuffer);
cudaGLMapBufferObject ((void**)&my_pointer, pixelBuffer);
my_kernel<<<64, 64>>>(my_pointer);
cudaGLUnmapBufferObject(pixelBuffer);
cudaGLUnregisterBufferObject(pixelBuffer);
glBindTexture(GL_TEXTURE_2D, tex_id);
glTexImage2D (..., 0, ..., w, h, 0, ..., (GLvoid*)0);
```



OpenGL 3.0 interop





OpenGL 3.0 interop

```
GLuint tex_id;
glGenTextures(1, & tex_id);
struct cudaGraphicsResource *tex_resource;
cudaGraphicsGLRegisterImage (&tex_resource, tex_id,
               GL_TEXTURE_2D, cudaGraphicsMapFlagsReadOnly);
cudaGraphicsMapResources (1, & tex_resource, 0);
cudaArray *array_ptr;
cudaGraphicsSubResourceGetMappedArray (&array_ptr, tex_resource, 0, 0);
cudaMemcpyToArray(array_ptr,0, 0,dest,size,cudaMemcpyDeviceToDevice);
cudaGraphicsUnmapResources(1, & tex_resource, 0);
```



Ray Tracing

#Фотореалистичный синтез изображений





#POV-Ray



Ray Tracing

#Фотореалистичный синтез изображений



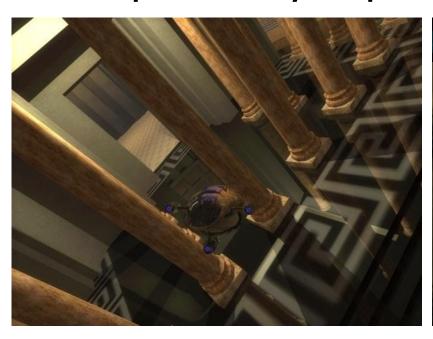


#POV-Ray



Real Time Ray Tracing

ж Скорость в ущерб качеству







Ray Tracing

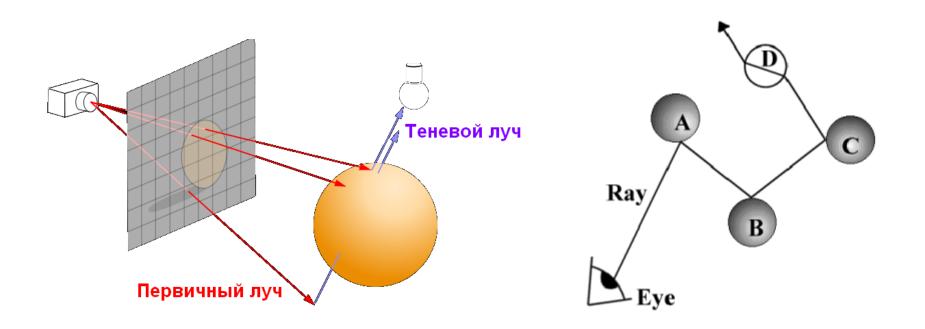
- **Ж** Точность
 - Path tracing
 - Фотонные карты
 - Распределенная трассировка лучей (стохастическая)
- **ж** Скорость
 - Обратная трассировка лучей
 - Растеризация + обратная трассировка лучей

Обратная трассировка лучей



Ж Алгоритм

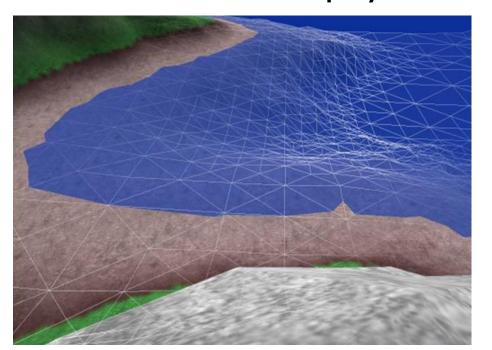
Первичные, теневые, отраженные лучи

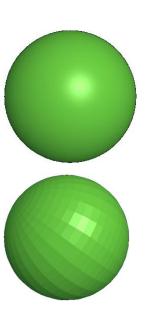


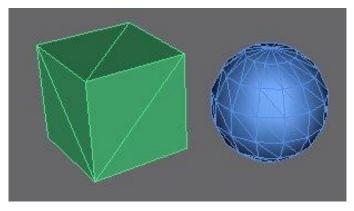


Ray Tracing

- **Ж** Представление 3D объектов
 - Аналитическое
 - Меши из треугольников





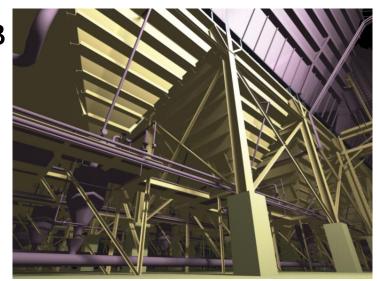




Ray Tracing

- **Ж** Поверхность задана как массив треугольников
- Узкое место поиск пересечения луча с поверхностью
 - 1 000 000 треугольников
 - № 1 000 000 лучей

 - $\triangle \log(N)^k * 10^6 (k \sim [1..2])$





ЖПростой вариант

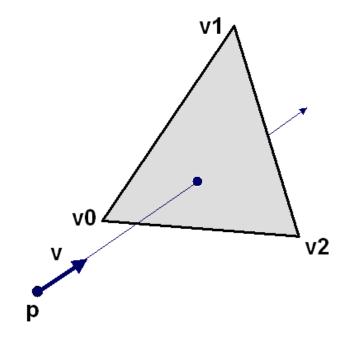
$$\triangle$$
 $Ax + By + Cz + D = 0$

Майти *t*

$$\times x = p.x + v.x*t$$

$$\times$$
 $y = p.y + v.y*t$

$$\times$$
 $z = p.z + v.z*t$



$$t = -\frac{(A*p.x + B*p.y + C*p.z + D)}{A*v.x + B*v.y + C*v.z}$$



ЖПростой вариант

 \triangle t известно

$$\times$$
 $z = p + v*t$

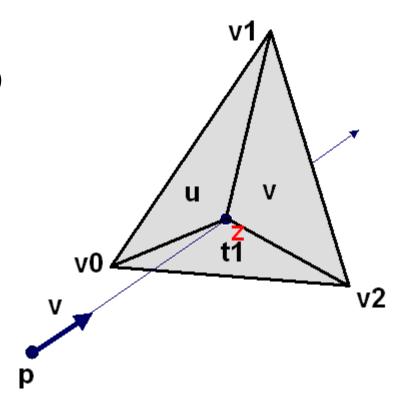
$$\boxtimes$$
 $S = cross(v1-v0, v2-v0)$

$$\boxtimes$$
 $u = cross(v1-z, v0-z)$

$$\times$$
 $v = cross(v1-z, v2-z)$

$$\times$$
 t1 = cross(v2-z, v0-z)

$$|u + v + t1 - S| < \varepsilon$$





ЖОптимизированный вариант

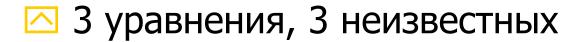
Барицентрические координаты

$$\boxtimes u := u/S, v := v/S, t1 := t1/S$$

$$\times t1 = 1 - u - v$$

$$z(u,v) = (1-u-v)*v1 + u*v2 + v*v0$$
$$z(t) = p + t*d$$

$$p + t * d = (1 - u - v) * v1 + u * v2 + v * v0$$

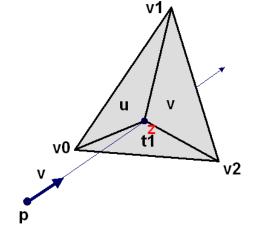




#Оптимизированный вариант

$$\begin{bmatrix} t \\ u \\ v \end{bmatrix} = \frac{1}{dot(P, E1)} * \begin{bmatrix} dot(Q, E2) \\ dot(P, T) \\ dot(Q, D) \end{bmatrix}$$





$$E1 = v1 - v0$$

$$E2 = v2 - v0$$

$$T = p - v0$$

$$P = cross(D, E2)$$

$$Q = cross(T, E1)$$

$$D = v$$

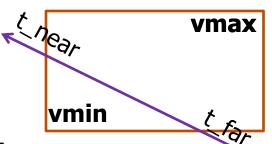


ЖПростой вариант

- Операции (* : 39, +/- : 53, / : 1)
- **Ж**Оптимизированный вариант
 - Операции (* : 23, +/- : 24, / : 1)
- **ж** Как считали нижнюю оценку?
 - использование mad вместо mul и add
 - 4*(N_mul + |N_add N_mul|)



Другие примитивы



- \triangle (vmin.x + rInv.pos.x) * rInv.dir.x;
- 6 add и 6 mul == 12 mad, 48 тактов
- **ж** Сфера
 - \sim 13 mad + sqrtf == 52 + 32 = 84 такта
 - меньше ветвлений
 - Иерархия из сфер не лучше иерархии из боксов

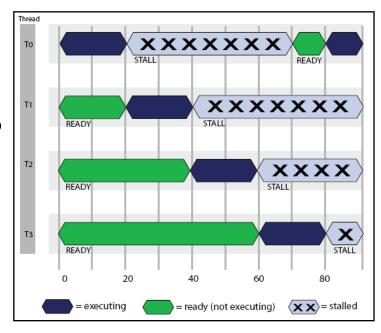


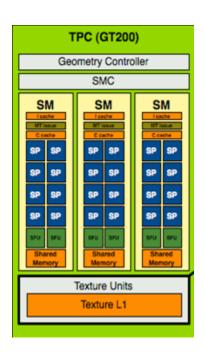


******Очень быстрое переключение

ЖВлияет:

- Регистры
- shared память
- Размер блока





$$occupancy = \frac{active warps}{max warps}$$

Occupancy

Регистры

active warps occupancy =





- Блоки по 8х8 нитей
- 128 регистров на нить

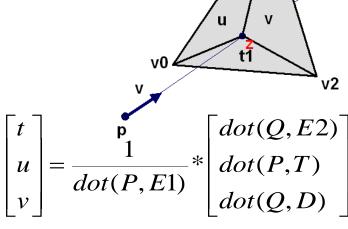


- рег <= 40: 3 блока, 6 warp-ов активны
- рег <= 32: 4 блока, 8 warp-ов активны
- рег <= 24: 5 блоков, 10 warp-ов активны
- рег <= 20: 6 блоков, 12 warp-ов активны
- рег <= 16: 8 блоков, 16 warp-ов активны



Регистры

- 6 регистров на луч
- 9 регистров на вершины
- 3 регистра на (t, u, v)
- 1 регистр на triNum
- 1 на счетчик в цикле
- 1 как минимум на tid
- 2 на min_t и min_id
- **ж** 23 уже занято!



$$E1 = v1 - v0$$

$$E2 = v2 - v0$$

$$T = p - v0$$

$$P = cross(D, E2)$$

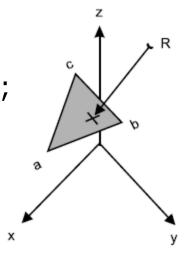
$$Q = cross(T, E1)$$

$$D = v$$

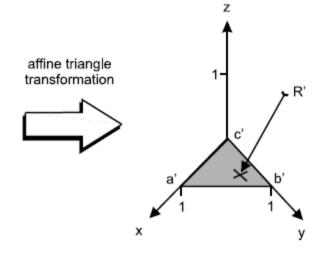


#Unit test

float3 o = mul3x4(m, origin); float3 d = $\frac{\text{mul3x3}}{\text{mul3x3}}$ (m, dir); float t = -o.z/d.z; float u = o.x + t*d.x; float v = o.y + t*d.y;



(a) world coordinate space



(b) unit triangle space

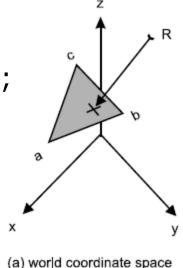
$$T_{\Delta}^{-1} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = A \quad T_{\Delta}^{-1} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = B$$

$$T_{\Delta}^{-1} \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = C \quad T_{\Delta}^{-1} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = N$$



#Unit test

float3 o = mul3x4(m, origin); float3 d = $\underline{\text{mul3x3}}$ (m, dir); float t = -o.z/d.z; float u = o.x + t*d.x; float v = o.y + t*d.y;



affine triangle transformation

A c' C' C' C' Y

(=,

(b) unit triangle space

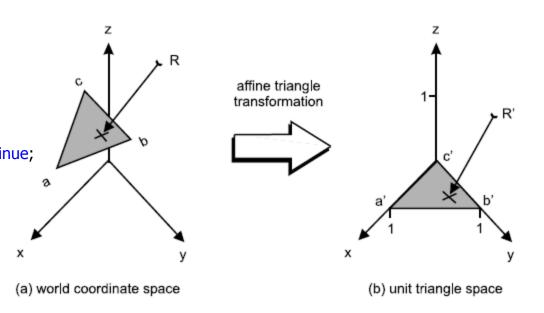
<u>Где m это T∆</u>

$$T_{\Delta}^{-1}(X) = \begin{pmatrix} A_x - C_x & B_x - C_x & N_x - C_x \\ A_y - C_y & B_y - C_y & N_y - C_y \\ A_z - C_z & B_z - C_z & N_z - C_z \end{pmatrix} \cdot X + \begin{pmatrix} C_x \\ C_y \\ C_z \end{pmatrix}$$



#Unit test

```
float4 row2 = tex1Dfetch(tex, triAddress+0);
(oz,dz) = ...
float t = -oz/dz;
if (t <= EPSILON E8M || t >= tTriangle) continue;
float4 row0 = tex1Dfetch(tex, triAddress+1);
(ox, dx) = ...
float u = ox + t*dx;
if (u < 0.f) continue;
float4 row1 = tex1Dfetch(tex, triAddress+2);
(oy dy) = ...
float v = oy + t*dy;
if(v >= 0 \&\& (u+v) <= 1.f)  {
    tTriangle = t;
    tIndex = currIndex;
 }
```



Операции (*: 20, +: 20, /: 1)

- ◆116-196 тактов
- ▶21 регистр на весь kernel
- ▶22 + поиск в kd дереве
- →25 + persistent threads



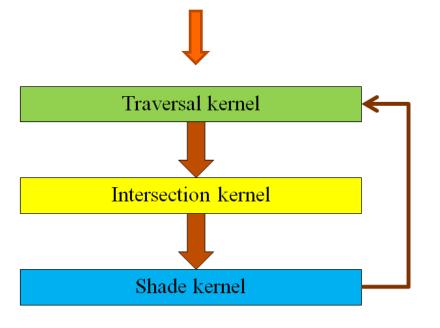
Экономия регистров

ЖДва подхода к экономии регистров

uber kernel



separate kernel



```
global void Uber kernel()
START:
  if (state == TRAVERSE)
    // ....
    if (node.leaf())
      // save nessesary data
      state = INTERSECTION;
      goto START;
  else if (state == INTERSECTION)
    // ...
  else if (state == SHADE)
    //...
```



Экономия регистров

Uber kernel

- **ж** Рекурсия (+)
- Ж Локальность данных (+)
- Ж Трудно профилировать (-)
- **Ж** Трудно оптимизировать (-)

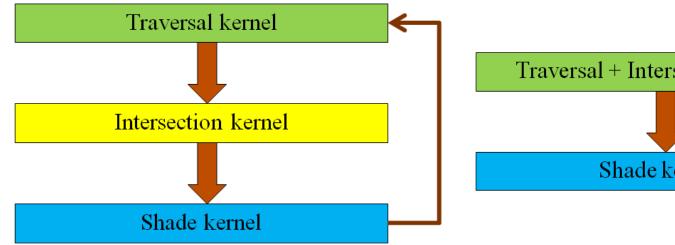
Separate kernel

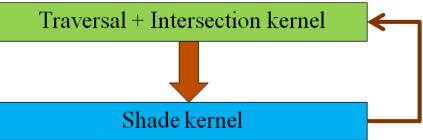
- Нет рекурсии (-)
- Обращения к DRAM (-)
- Остановка ядер (-)
- Легко профилировать и оптимизировать (+)
- Гибкость (+)
- Проще интегрировать с
 CPU реализациями (+)



Архитектура

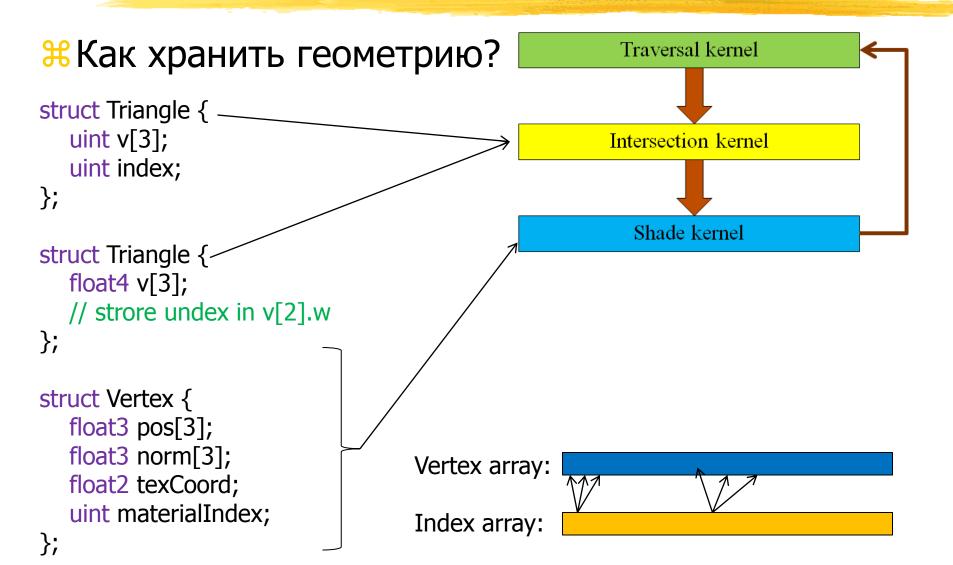
- Ядро пересечений 24-32 регистра
- Но это только пересечения
- Нужно разбить алгоритм трассировки на несколько ядер







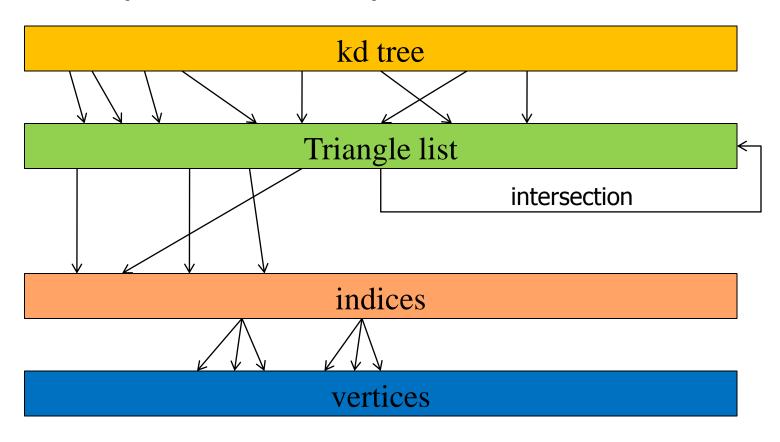
Архитектура





Архитектура

Ж Как хранить геометрию?

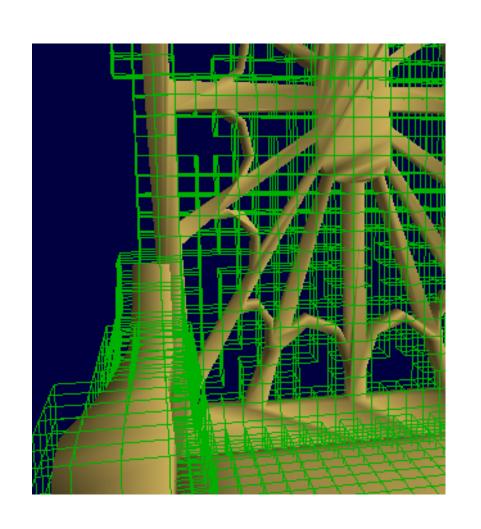


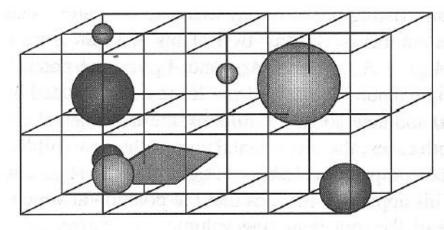


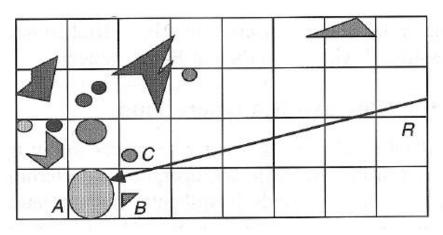




Регулярная сетка





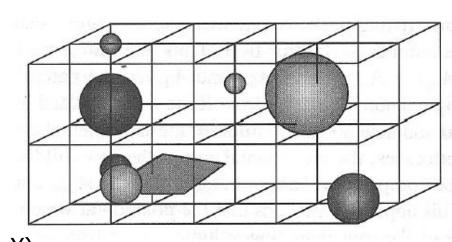


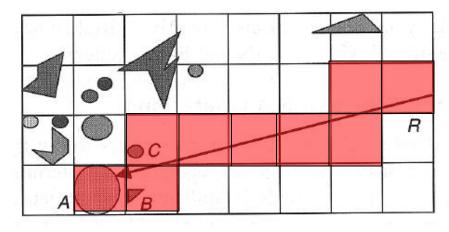


Регулярная сетка

ЖРегулярная сетка

```
if (tMaxX <= tMaxY && tMaxX <= tMaxZ)</pre>
  tMaxX += tDeltaX;
  x += stepX;
else if (tMaxY <= tMaxZ && tMaxY <= tMaxX)
  tMaxY += tDeltaY;
  y += stepY;
else
  tMaxZ += tDeltaZ;
  z += stepZ;
```

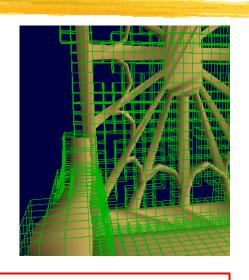






Регулярная сетка

- **#** Преимущества
 - Просто и быстро строится
 - Простой алгоритм траверса
- **ж** Недостатки
 - Плохо справляется с пустым пространством
 - Требует много памяти
 - Много повторных пересечений –отвратительно разбивает геометрию
- **Ж** Только для небольших сцен (1-50K)

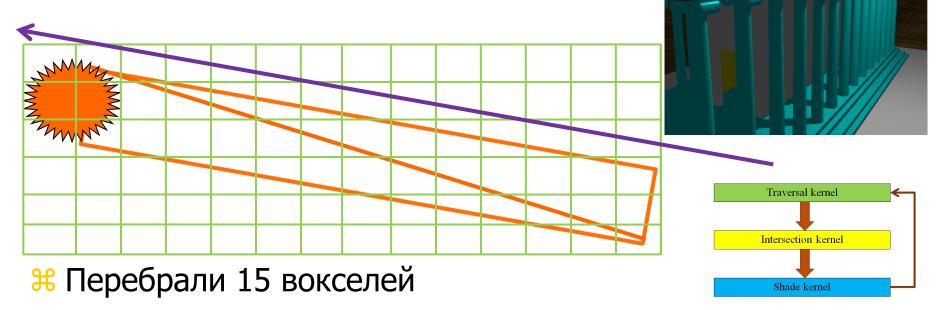




Регулярная сетка

ЖПочему сетка плохо разбивает

геометрию?

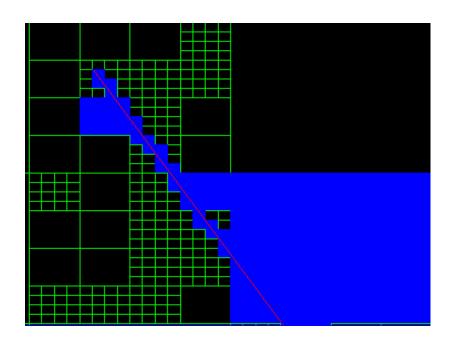


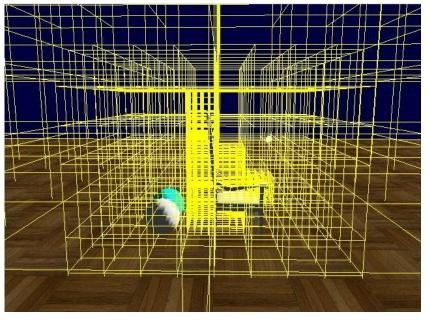
7 раз посчитали пересечение с одним и тем же треугольником!



Иерархическая сетка

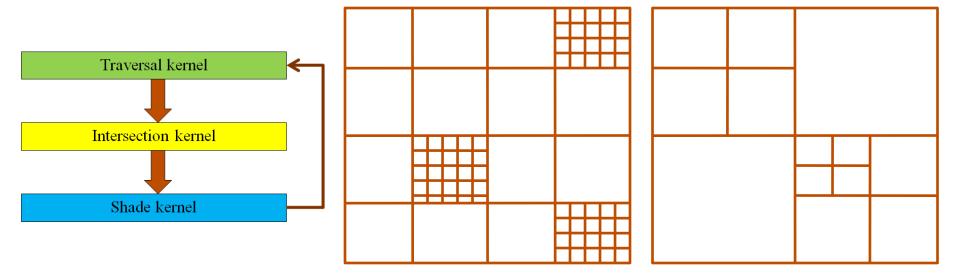
- **Ж**Небольшое число вокселей
- жРекурсивно разбиваем воксели в местах с плотной геометрией





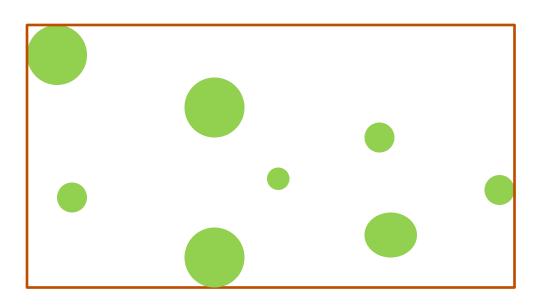
Что дает иерархическая сетка?

- + Решает проблему чайника на стадионе
- Переход между узлами вычислительно сложен
- + 12 регистров как минимум
- Нужно устранять рекурсию



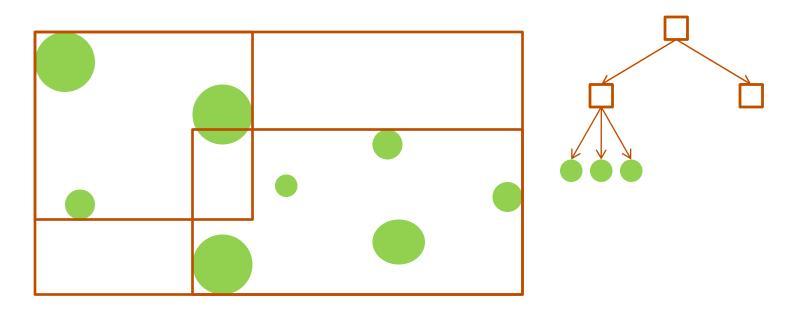


#Bounding Volume Hierarchy



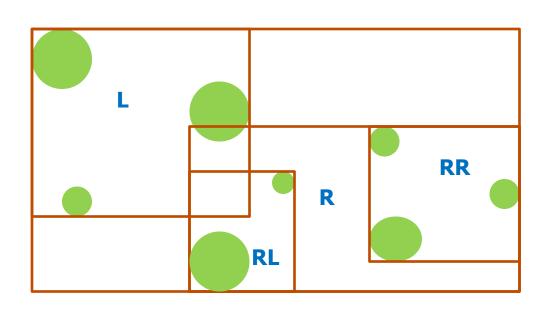


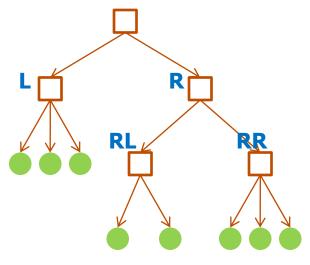
****Bounding Volume Hierarchy**





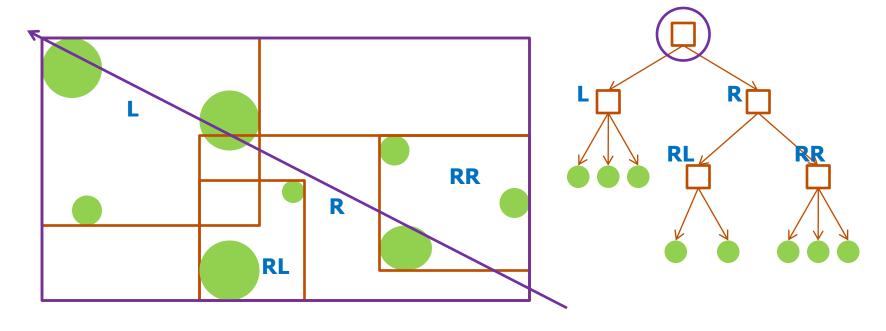
****Bounding Volume Hierarchy**





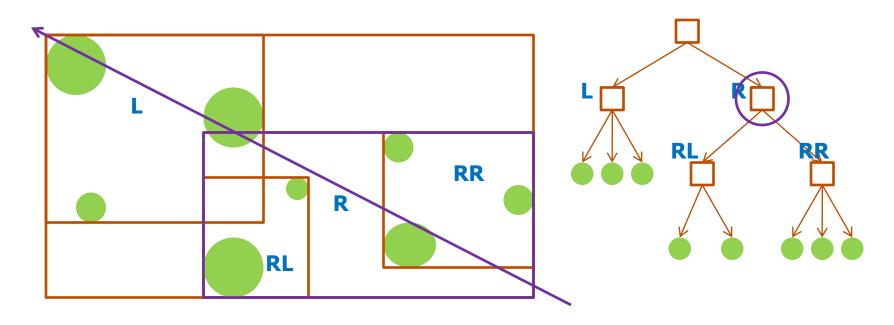


ж Траверс на CPU





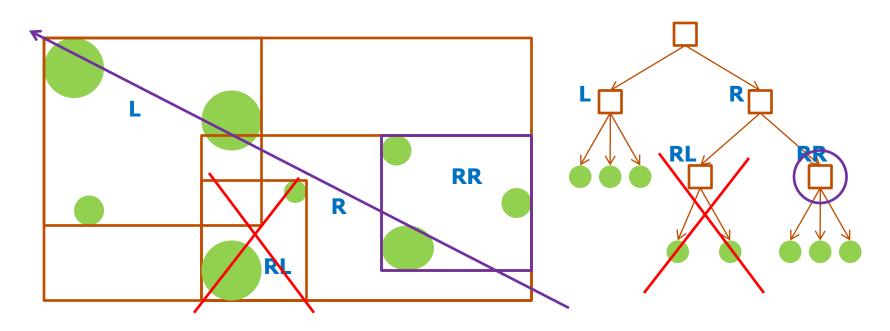
ж Траверс на CPU



Стек: L



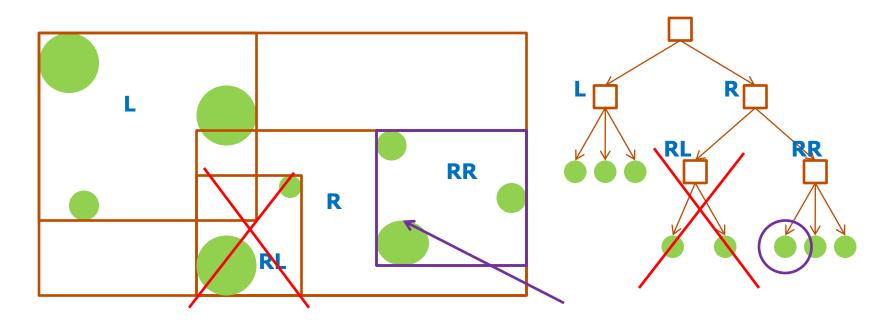
ж Траверс на CPU



Стек: L

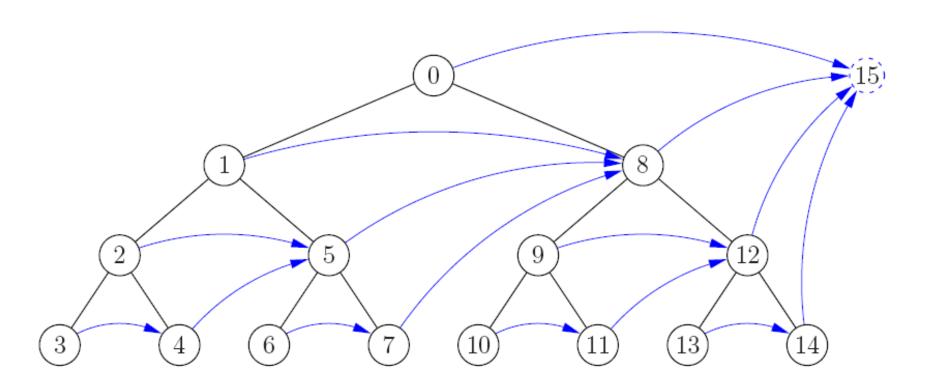


ж Траверс на CPU

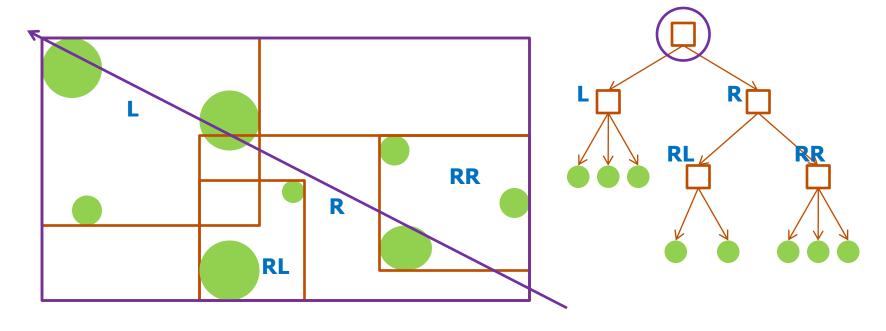


Стек: L

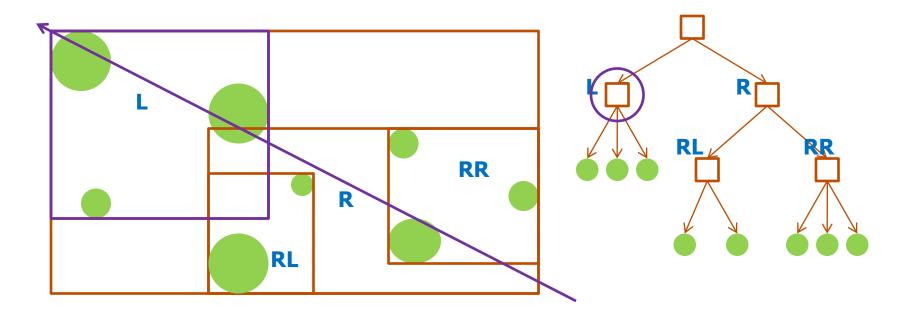




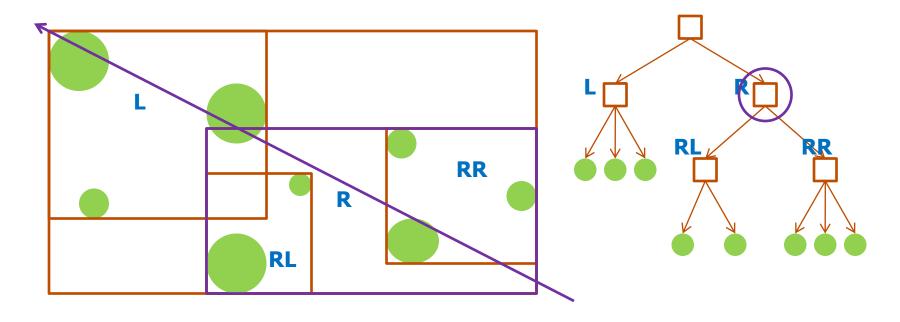




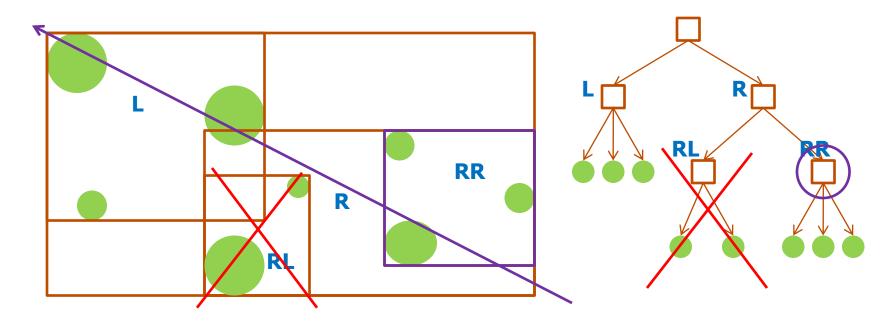




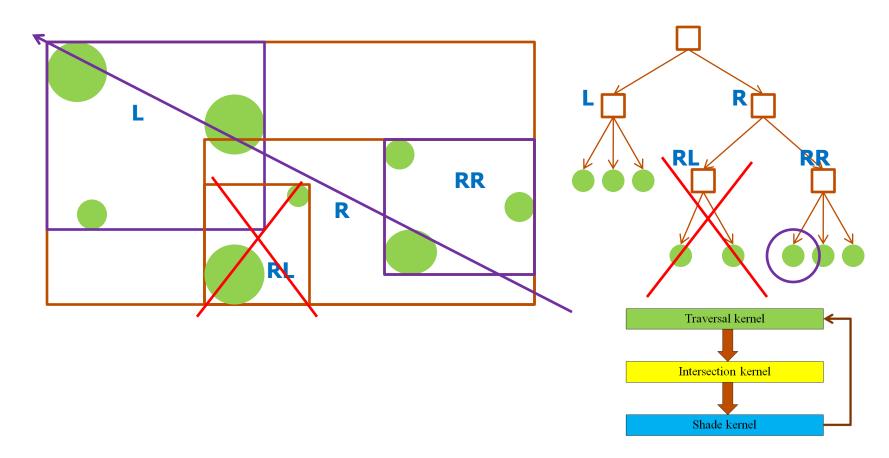








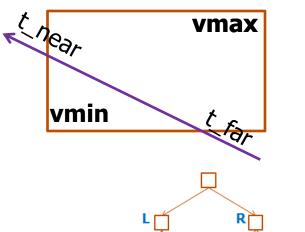






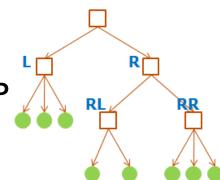
- - № Бокс 6 регистров

 - nodeOffset, leftOffset, tid 3
- **Ж** Пересечение луча с боксом
 - Минимум по всем 6 плоскостям
 ⋉ (vmin[0] + rInv.pos[0]) * rInv.dir[0];





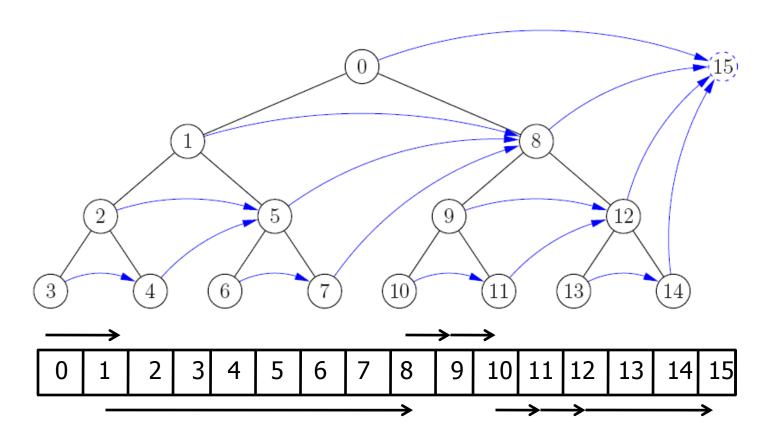
- - 24 mad-а покрывают латентность текстурной памяти



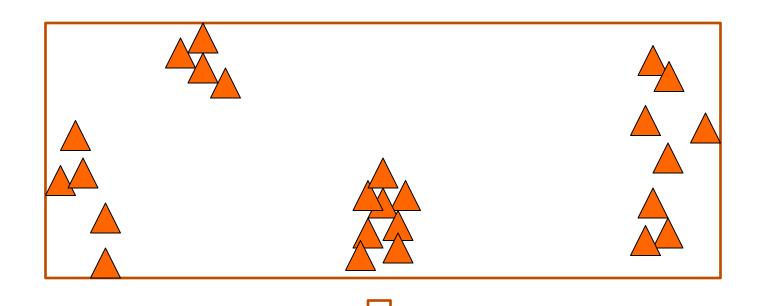
- 1. Стек на локальной памяти
 - Локальная память это не так медленно, как может показаться
- 2. Бесстековый алгоритм
 - Перебираем массив всегда строго слева направо



Безстековый траверс на GPU

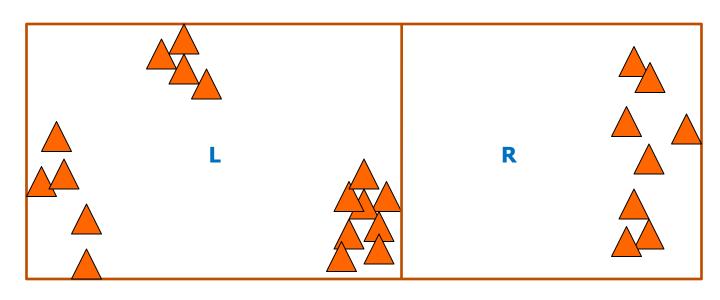




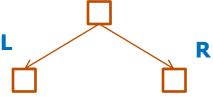


```
struct KdTreeNode
{
    float split;
    uint leftOffset: 29;
    uint splitAxis: 2;
    uint leaf: 1;
};
```

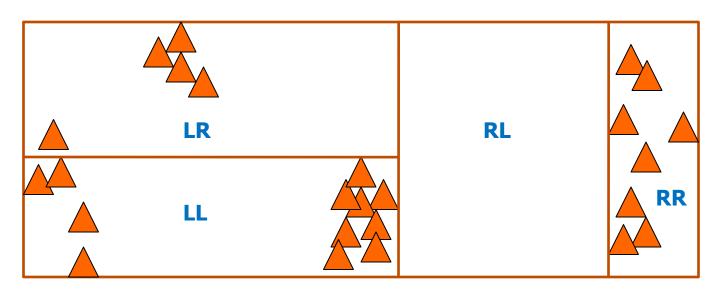




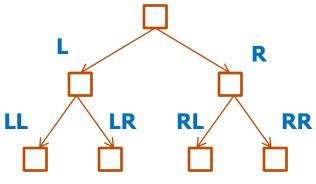
```
struct KdTreeNode
{
    float split;
    uint leftOffset: 29;
    uint splitAxis: 2;
    uint leaf: 1;
};
```



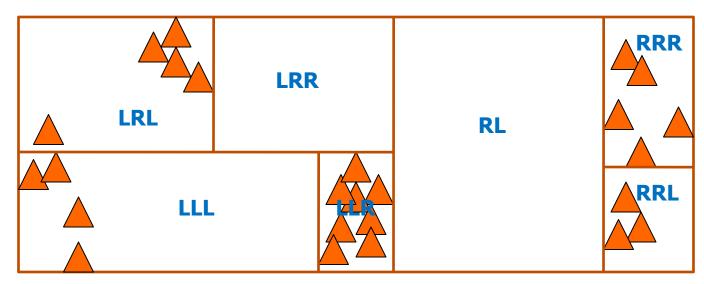




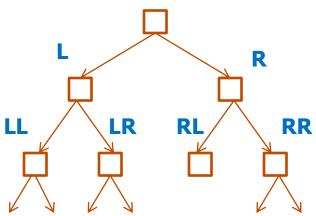
```
struct KdTreeNode
{
    float split;
    uint leftOffset: 29;
    uint splitAxis: 2;
    uint leaf: 1;
};
```



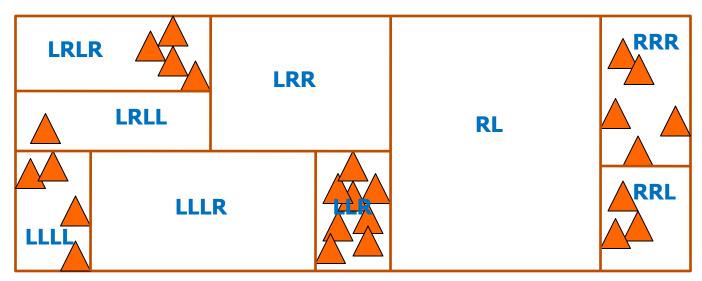




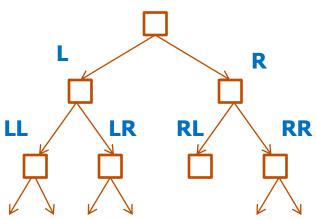
```
struct KdTreeNode
{
    float split;
    uint leftOffset: 29;
    uint splitAxis: 2;
    uint leaf: 1;
};
```



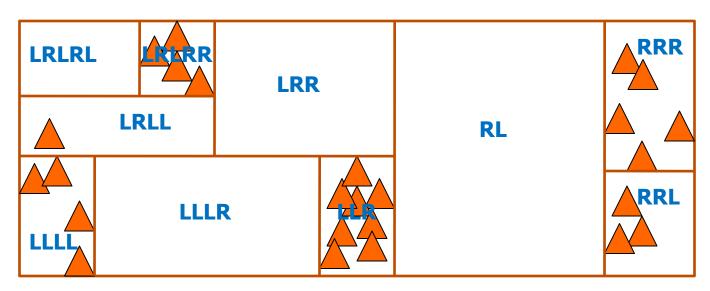




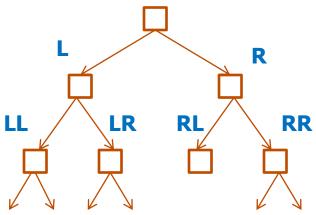
```
struct KdTreeNode
{
    float split;
    uint leftOffset: 29;
    uint splitAxis: 2;
    uint leaf: 1;
};
```







```
struct KdTreeNode
{
    float split;
    uint leftOffset: 29;
    uint splitAxis: 2;
    uint leaf: 1;
};
```

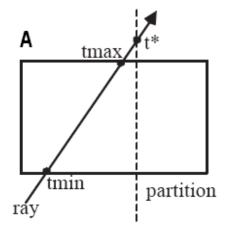


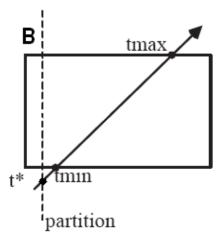


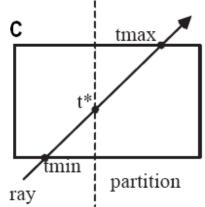
- **Ж**Алгоритм траверса
- **Ж**Регистры − 13 min:

 - node − 2

 - На практике удалось уложиться в 16!
 - Стек: локальная память

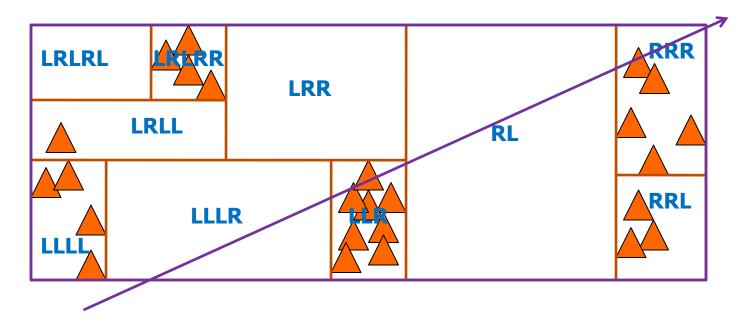








ЖАлгоритм траверса

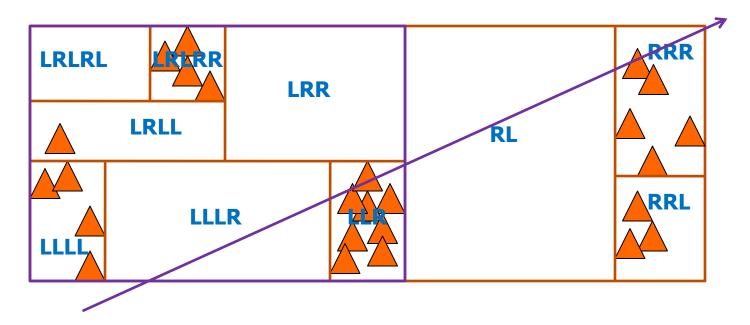


Стек:

Текущий узел:



ЖАлгоритм траверса

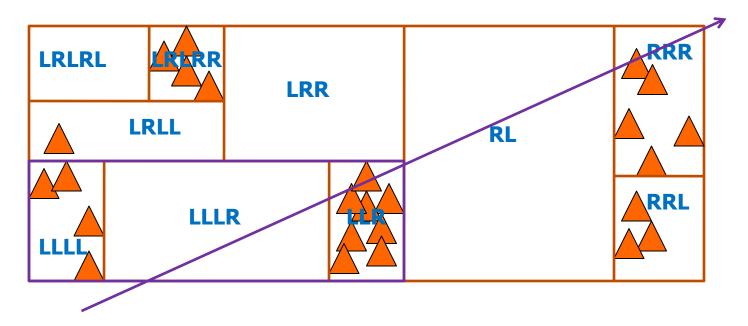


Стек: R

Текущий узел: L



ЖАлгоритм траверса

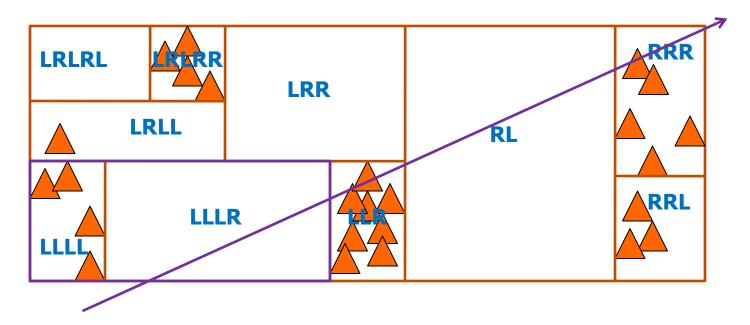


Стек: R

Текущий узел: LL



ЖАлгоритм траверса

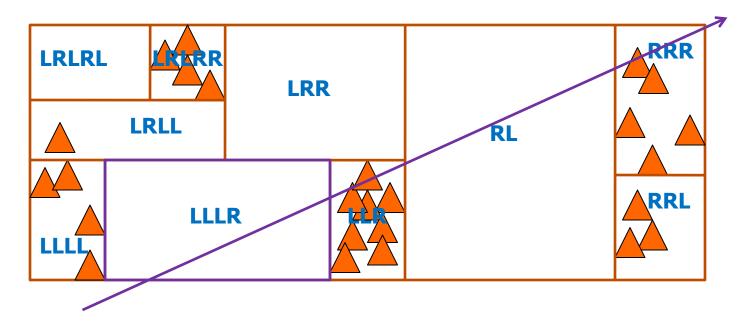


Стек: LLR, R

Текущий узел: LLL



ЖАлгоритм траверса

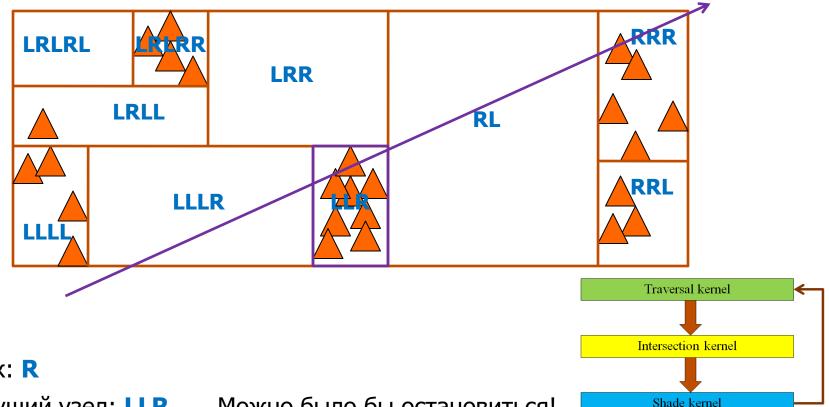


– Стек: LLR, R

Текущий узел: LLLR



ЖАлгоритм траверса

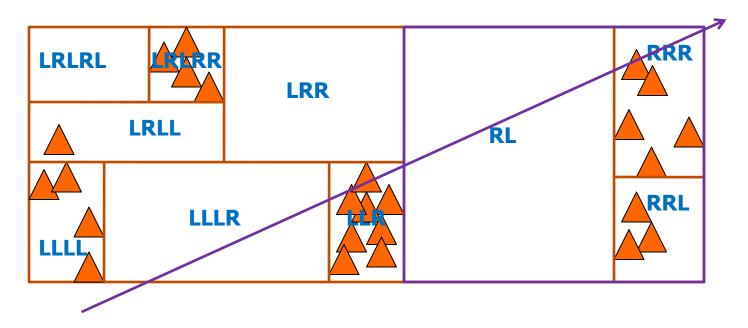


Стек: R

Текущий узел: LLR Можно было бы остановиться!



ЖАлгоритм траверса

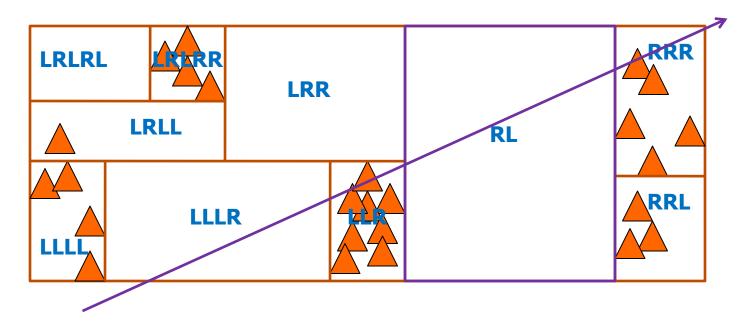


Стек:

Текущий узел: R



ЖАлгоритм траверса

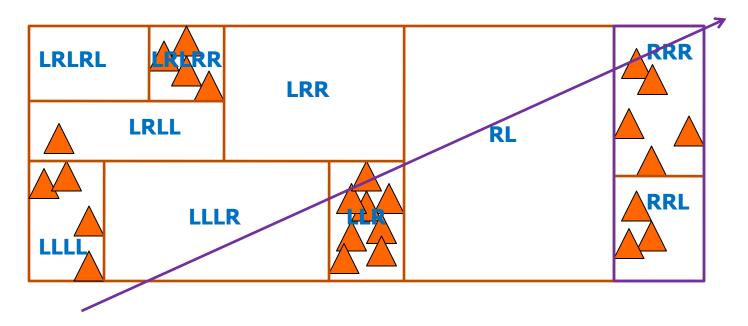


- Стек: **RR**

Текущий узел: RL



ЖАлгоритм траверса

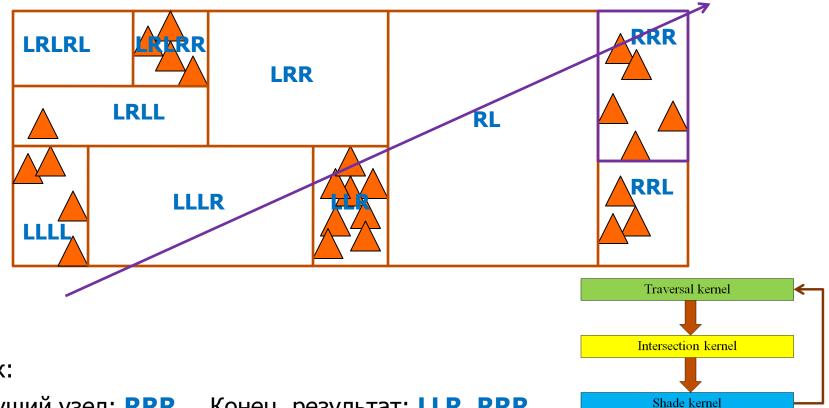


Стек:

Текущий узел: RR



ЖАлгоритм траверса



Стек:

Текущий узел: RRR Конец, результат: LLR, RRR



Производителность

Persistent threads

```
__global___ void my_kernel()
{
    for (int warp = 0; int warp < 16; warp++)
    {
        int tid = blockDim.x*blockIdx.x + threadIdx.x + warp*step;
        ... // тут код вашего кёрнела
    }
}
```



Производителность

- **Ж** Conference Room (281К треугольников)
- ₩~40М лучей в секунду на GTX260

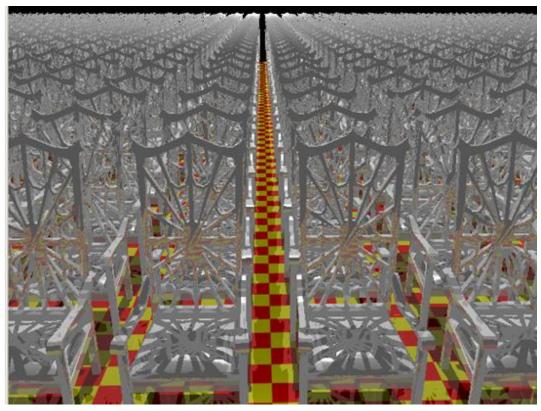






Производителность

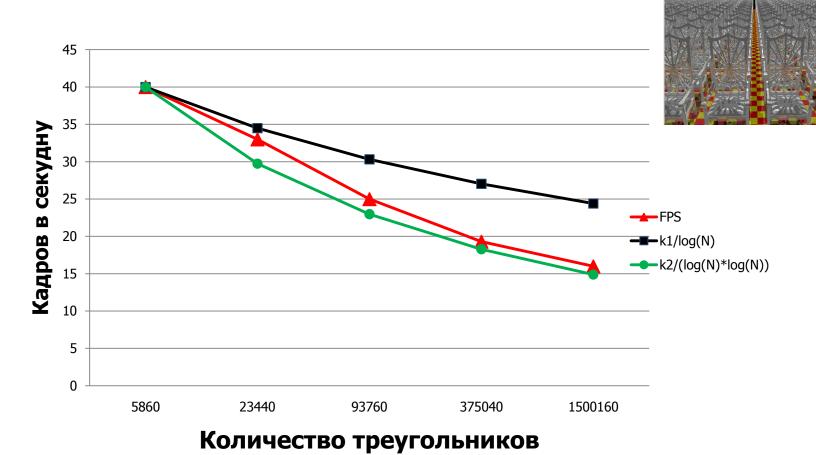
Ж Стулья (1.5М треугольников)





Производительность







kd-tree vs BVH на CUDA

- **Ж** BVH со стеком на локальной памяти
 - Покрывается латентность текстурной памяти
 - Меньше глубина
 - Лишние плоскости
- ★ kd-tree
 - Экономит регистры
 - Можно эффективнее задействовать кэш?
 - 1 mad и пара ветвлений на одну tex1Dfetch

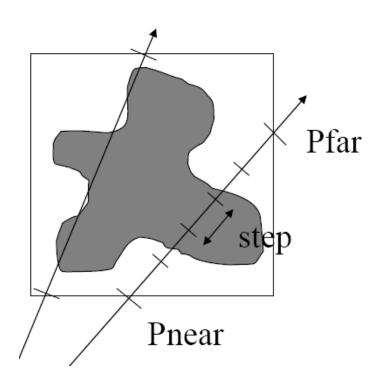


Резюме

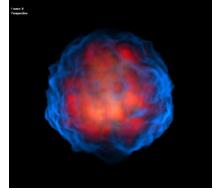
- **#OpenGL** interop
 - текстуры, PBO, VBO, FBO
- **#Occupancy**
 - Оптимизации по локальным переменным
 - Uber kernel
 - Separate kernel
- **#Persistent threads**

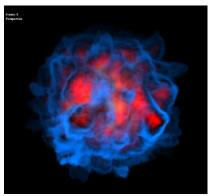


Ray marching



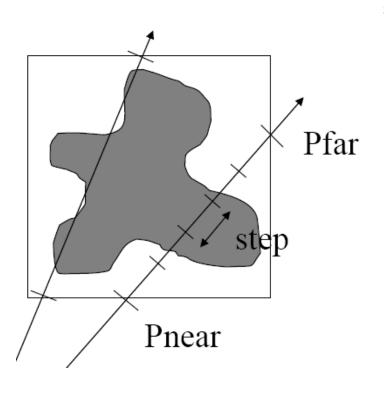








Ray marching



```
float4 RayMarchPS(Ray eyeray : TEXCOORDO,
                  uniform int steps) : COLOR
    eyeray.d = normalize(eyeray.d);
    // calculate ray intersection with bounding box
    float tnear, tfar;
    bool hit = IntersectBox(eyeray, boxMin, boxMax, tnear, tfar);
    if (!hit) discard;
    if (tnear < 0.0) tnear = 0.0;
    // calculate intersection points
    float3 Pnear = eyeray.o + eyeray.d*tnear;
    float3 Pfar = eyeray.o + eyeray.d*tfar;
    // march along ray, accumulating color
    half4 c = 0;
    half3 step = (Pnear - Pfar) / (steps-1);
    half3 P = Pfar;
    for(int i=0; i<steps; i++) {
            half4 s = VOLUMEFUNC(P);
            c = s.a*s + (1.0-s.a)*c;
            P += step;
    c /= steps;
    return c;
```



