# ОСОБЕННОСТИ РЕАЛИЗАЦИИ АЛГОРИТМОВ ТРАССИРОВКИ ЛУЧЕЙ НА CUDA

Докладчик:

Фролов В.А. (ВМиК МГУ)

Научный руководитель:

Игнатенко А.В. (ВМиК МГУ)

Лекторы:

Боресков А.В. (ВМиК МГУ)

Харламов A.A. (NVidia)

#### План

- Ж Метод грубой силы
  - CUDA
- **ж** Ускоряющие структуры
  - Регулярные и иерархические сетки
  - BVH
  - kd деревья

## **Ray Tracing**

#### **#**Фотореалистичный синтез изображений





**#POV-Ray** 

## **Ray Tracing**

#### **#**Фотореалистичный синтез изображений





**#POV-Ray** 

### **Real Time Ray Tracing**

#### **ж** Скорость в ущерб качеству





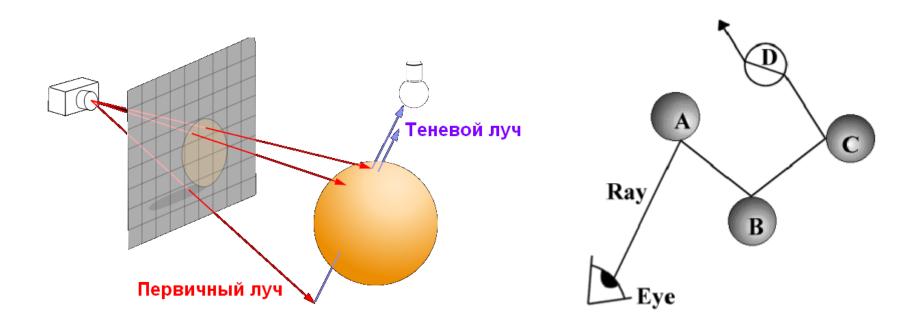
## **Ray Tracing**

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

# Обратная трассировка лучей

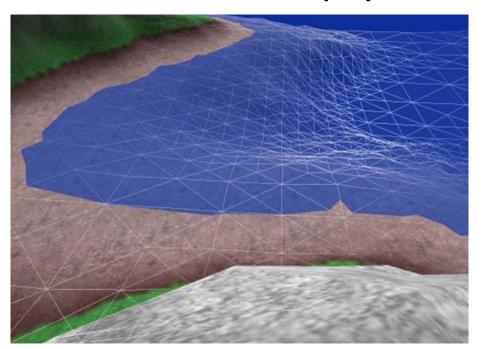
#### **Ж** Алгоритм

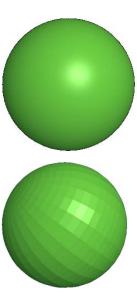
Первичные, теневые, отраженные лучи

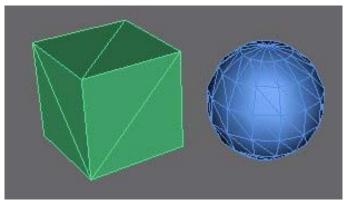


## **Ray Tracing**

- # Представление 3D объектов
  - Аналитическое
  - Меши из треугольников



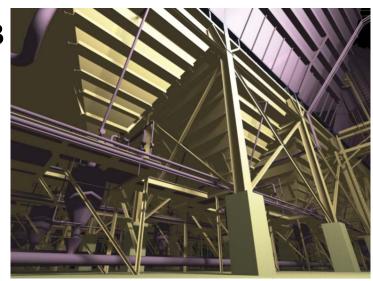




### **Ray Tracing**

- **Ж** Поверхность задана как массив треугольников
- Эзкое место поиск пересечения луча с поверхностью
  - 1 000 000 треугольников
  - 🔼 1 000 000 лучей

  - $\triangle$  (log(N))<sup>k\*</sup>10<sup>6</sup> (k~[1..2])



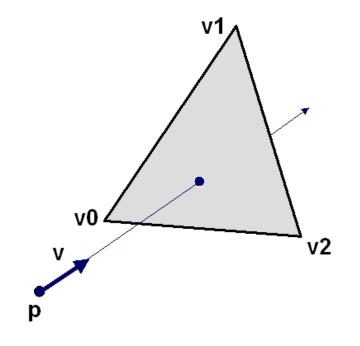
#### **Ж**Простой вариант

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

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

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

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



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

#### **Ж**Простой вариант

 $\triangle$  t известно

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

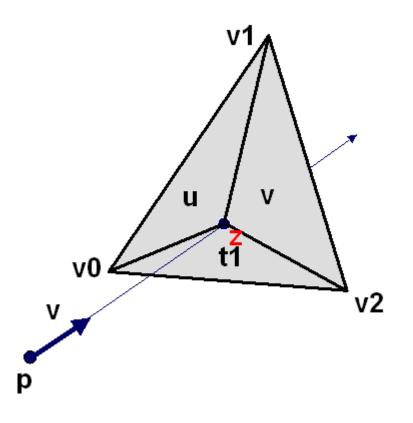
$$\boxtimes$$
  $S = cross(v1-v0, v2-v0)$ 

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

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

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

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



#### **Ж**Оптимизированный вариант

Барицентрические координаты

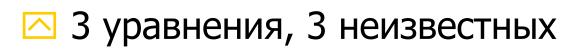
$$\boxtimes u := u/S, v := v/S, t1 := t1/S$$

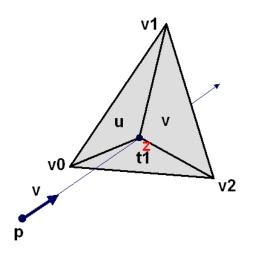
$$\times t1 = 1 - u - v$$

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

$$z(t) = p + t * d$$

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





#### **Ж**Оптимизированный вариант

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

$$E1 = v1 - v0$$

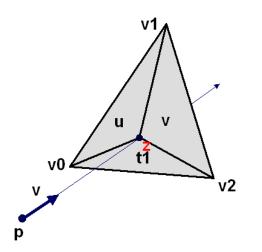
$$E2 = v2 - v0$$

$$T = p - v0$$

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

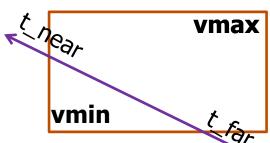
$$Q = cross(T, E1)$$

$$D = v$$



- **Ж**Простой вариант
  - Операции (\* : 39, +/- : 53, / : 1)
- **Ж**Оптимизированный вариант
  - Операции (\* : 23, +/- : 24, / : 1)
    - 🔀 132-224 такта
- **ж** Как считали нижнюю оценку?
  - использование mad вместо mul и add
  - 4\*(N\_mul + |N\_add N\_mul|)

#### Другие примитивы

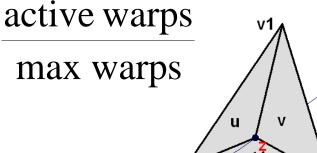


- (vmin.x + rInv.pos.x) \* rInv.dir.x;
- 6 add и 6 mul == 12 mad, 48 тактов
- **ж** Сфера
  - $\sim$  13 mad + sqrtf == 52 + 32 = 84 такта
  - меньше ветвлений
  - Иерархия из сфер не лучше иерархии из боксов

## **Multiprocessor Occupancy**

**#** Регистры

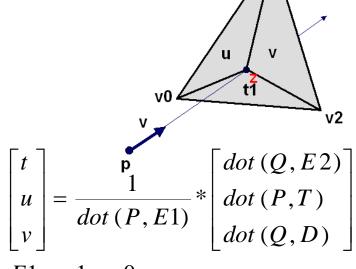
 $occupancy = \frac{1}{2}$ 



- № 8192 регистра на SM
- Блоки по 8х8 нитей
- 128 регистров на нить
  - пусс не дает столько регистров, почему?
  - 区 рег <= 40: 3 блока, 6 warp-ов активны
  - 区 рег <= 32: 4 блока, 8 warp-ов активны
  - рег <= 24: 5 блоков, 10 warp-ов активны
    </p>

#### **#** 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!



$$E1 = v1 - v0$$

$$E2 = v2 - v0$$

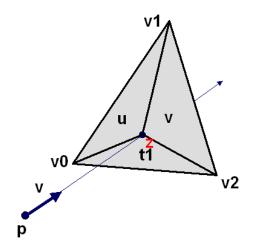
$$T = p - v0$$

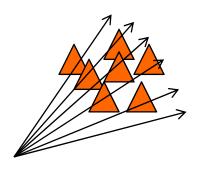
$$P = cross(D, E2)$$

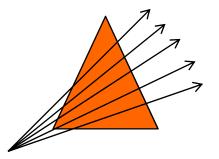
$$Q = cross(T, E1)$$

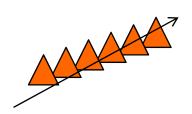
$$D = v$$

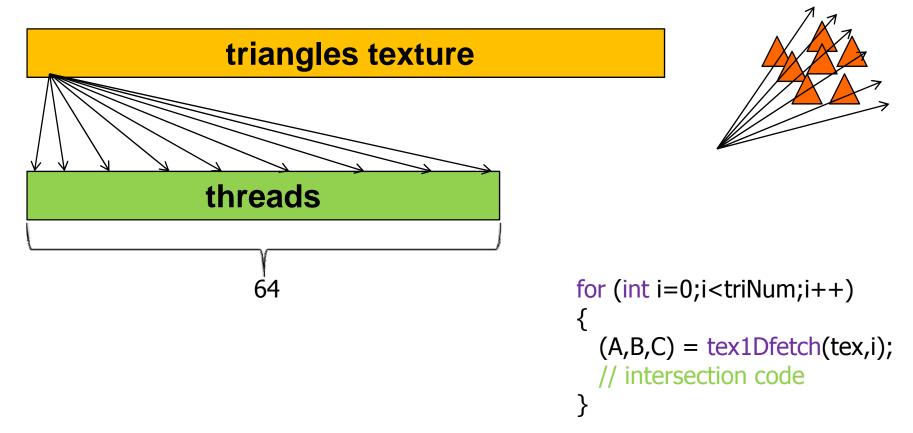
- ₩ Организация сетки блоки 8x8
- **ж** Что общее для нитей блока?
  - 1. Свой луч, свой треугольник
  - 2. Свой луч, общий треугольник
  - 3. Общий луч, свой треугольник

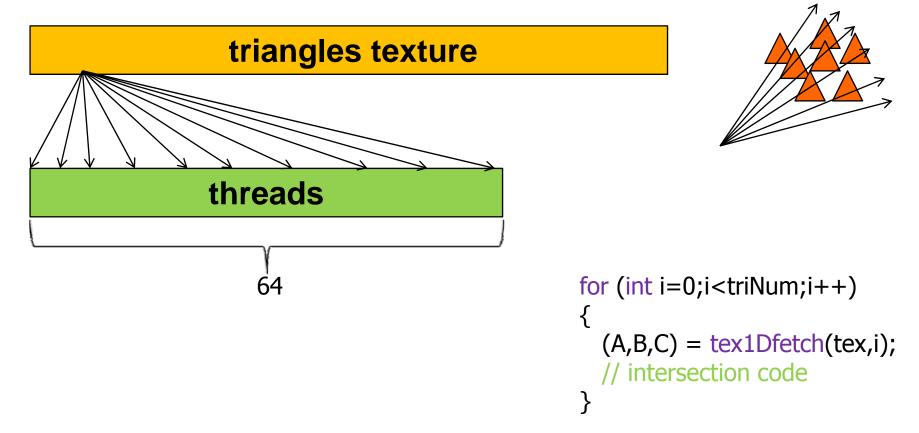


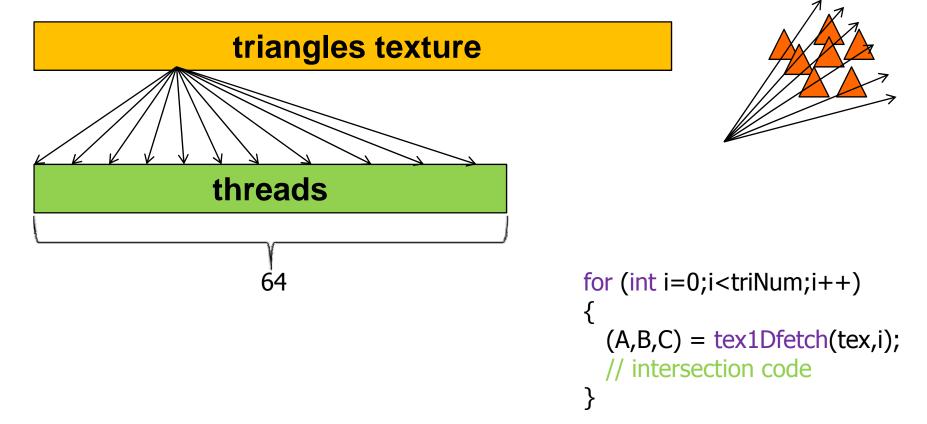


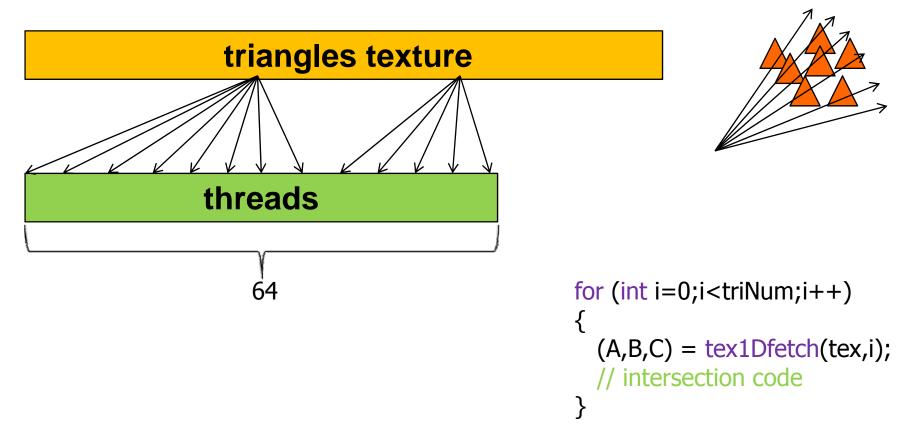


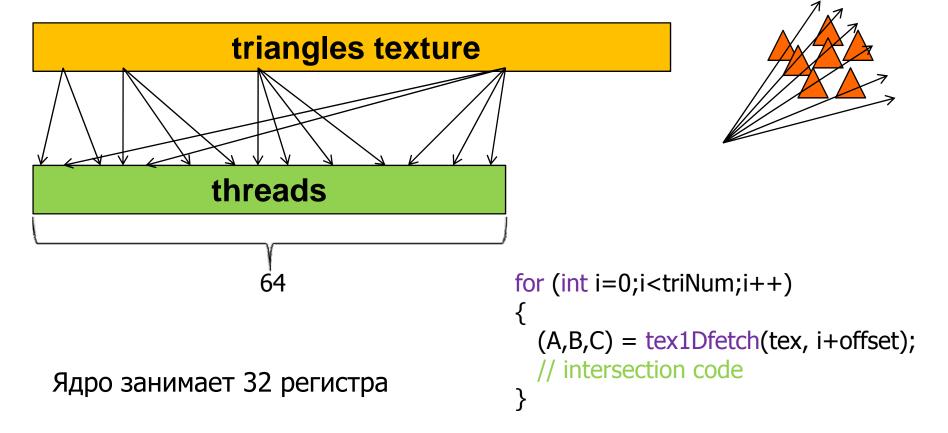


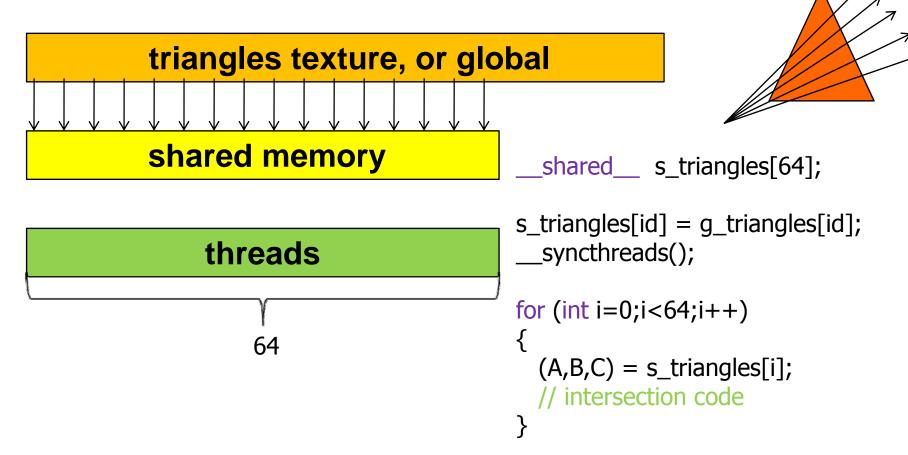






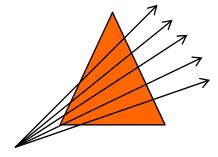


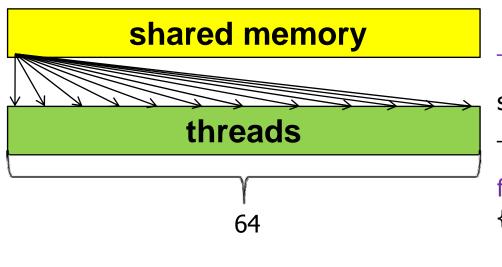




ЖСвой луч, общий треугольник

triangles texture, or global



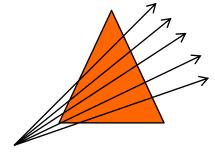


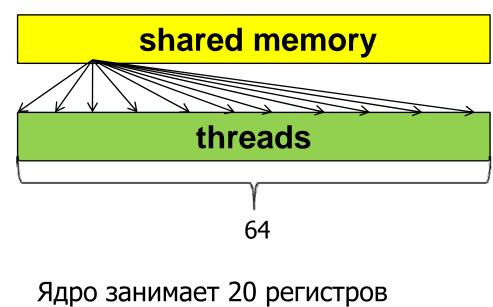
```
__shared__ s_triangles[64];
s_triangles[id] = g_triangles[id];
__syncthreads();

for (int i=0;i<64;i++)
{
    (A,B,C) = s_triangles[i];
    // intersection code
}</pre>
```

ЖСвой луч, общий треугольник

triangles texture, or global

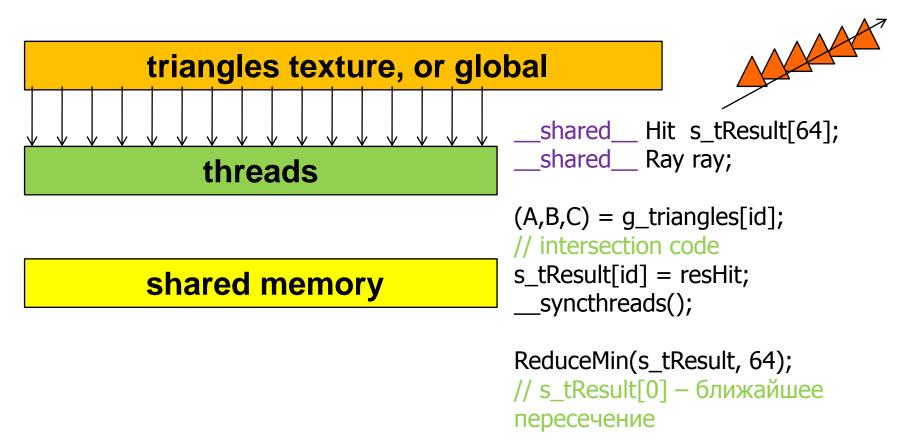




```
__shared__ s_triangles[64];
s_triangles[id] = g_triangles[id];
__syncthreads();

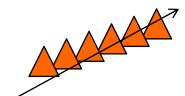
for (int i=0;i<64;i++)
{
    (A,B,C) = s_triangles[i];
    // intersection code
}</pre>
```

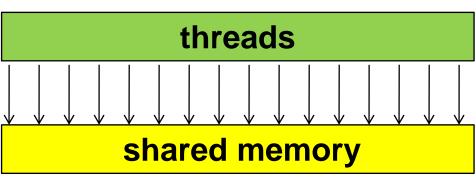
ЖОбщий луч, свой треугольник



ЖОбщий луч, свой треугольник

#### triangles texture, or global





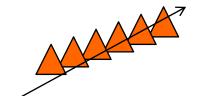
```
__shared__ Hit s_tResult[64];
__shared__ Ray ray;

(A,B,C) = g_triangles[id];
// intersection code
s_tResult[id] = resHit;
__syncthreads();

ReduceMin(s_tResult, 64);
// s_tResult[0] - ближайшее
пересечение
```

#### ЖОбщий луч, свой треугольник

#### triangles texture, or global



#### threads

shared memory

ReduceMin

Ядро занимает 21 регистр

```
__shared__ Hit s_tResult[64];
__shared__ Ray ray;
```

```
(A,B,C) = g_triangles[id];
// intersection code
s_tResult[id] = resHit;
__syncthreads();
```

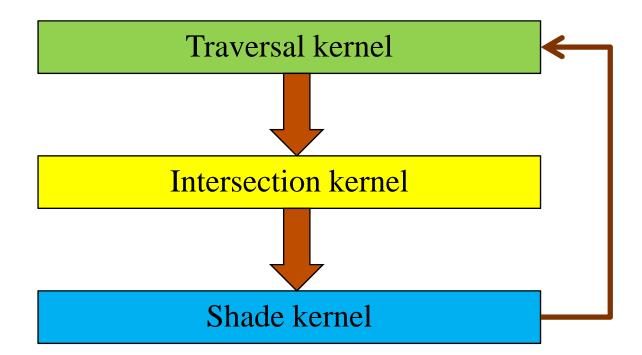
ReduceMin(s\_tResult, 64); // s\_tResult[0] – ближайшее пересечение

- Этанизация сетки блоки 8x8
- **ж** Что общее для нитей блока?
  - 1. Свой луч, свой треугольник (100%)
  - 2. Свой луч, общий треугольник (130%)
  - 3. Общий луч, свой треугольник (40-50%)



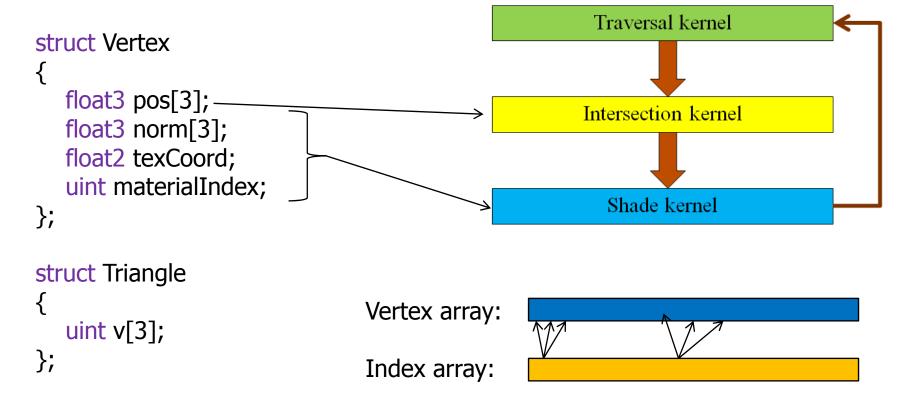
## Архитектура рейтрейсера

- Нужно разбить алгоритм трассировки на несколько ядер



### Архитектура рейтрейсера

- **Ж** Как хранить геометрию?



## Архитектура рейтрейсера

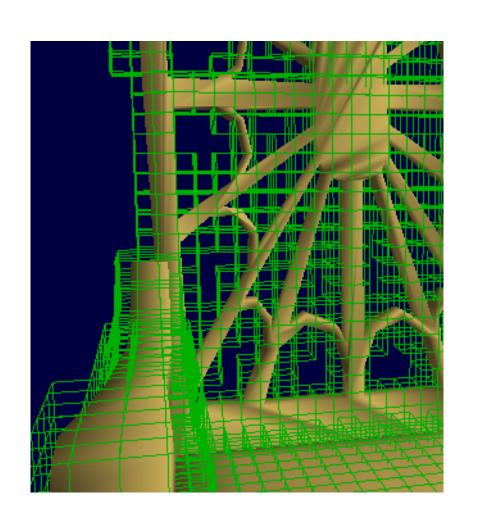
- **Ж** Дублирование геометрии

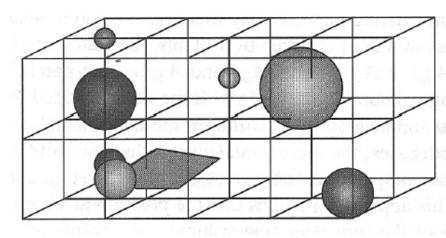
```
struct Triangle
  vec3f v[3];
  unsigned int selfIndex;
}; // 40
struct Sphere
  PackedSphere3f sph;
  unsigned int selfIndex;
  unsigned int dummy;
}; // 24
// выборки по float2
```

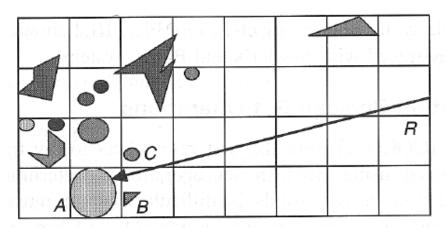
```
Traversal kernel
         Intersection kernel
            Shade kernel
struct Hit
   float t;
   unsigned int objectIdAndType;
};
```



## Регулярная сетка



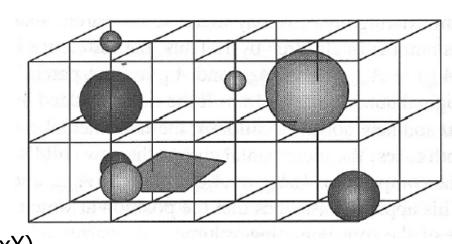


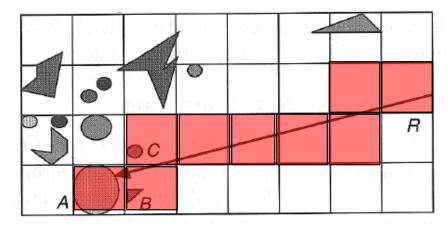


## Регулярная сетка

#### **Ж**Регулярная 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;
```





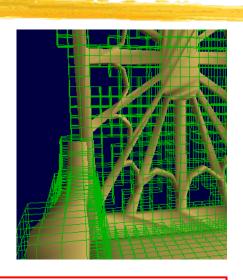
#### Регулярная сетка и CUDA

- **#** Преимущества
  - Занято как минимум 12 + 2 регистров
  - Z-ordering для 3D массива, кэш промахи↓
- **ж** Недостатки!
  - Жуткая нагрузка на текстурные блоки, латентность памяти не покрывается

вычислениями

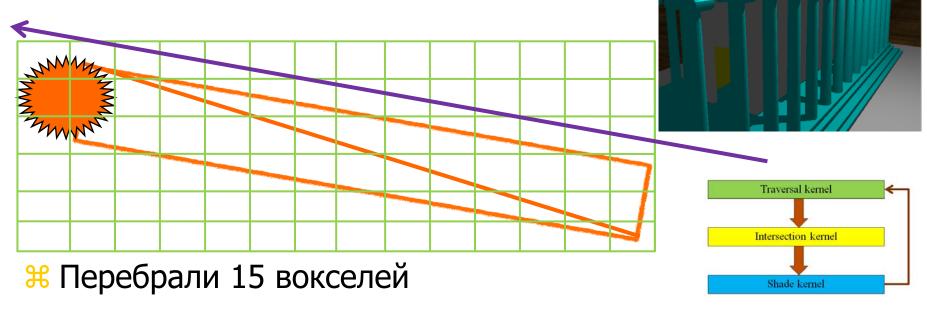
#### Регулярная сетка

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



#### Регулярная сетка

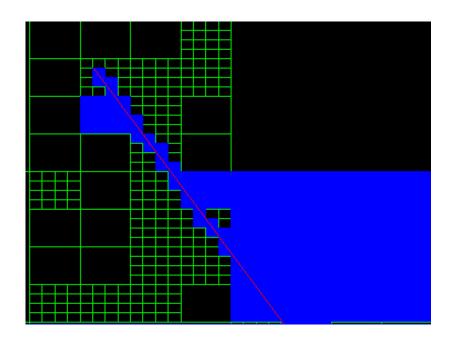
**ЖПочему сетка плохо разбивает геометрию?** 

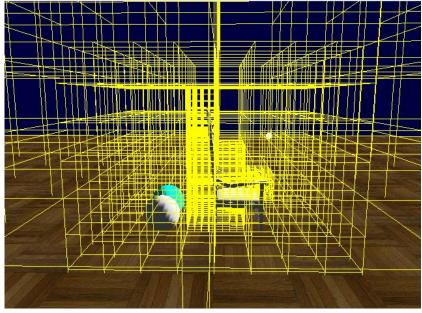


# 7 раз посчитали пересечение с одним и тем же треугольником!

#### Иерархическая сетка

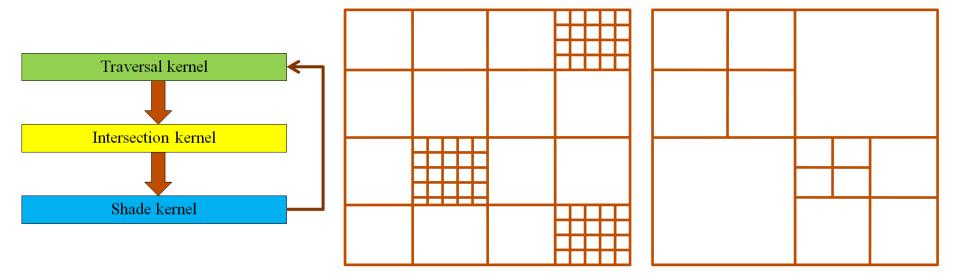
- **Ж**Небольшое число вокселей
- **ЖРекурсивно разбиваем воксели в местах с** плотной геометрией



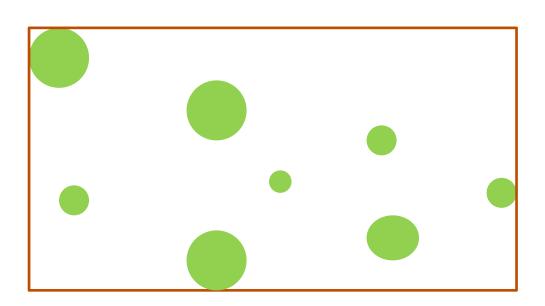


# Что дает иерархическая сетка?

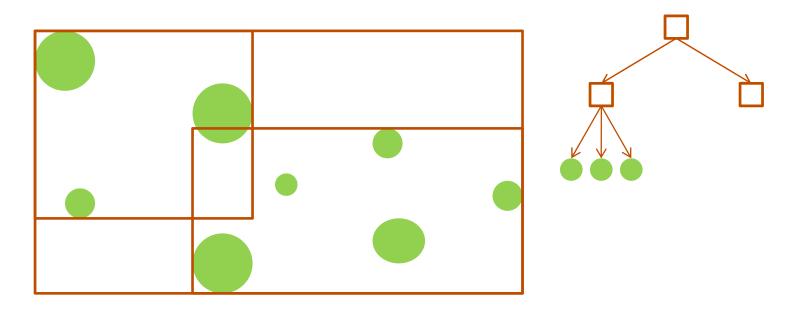
- + Решает проблему чайника на стадионе
- Переход между узлами вычислительно сложен
- + 12 регистров как минимум
- Нужно устранять рекурсию



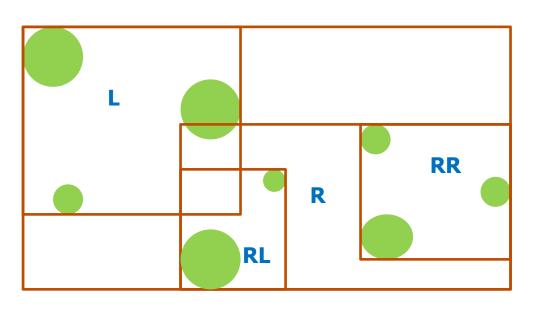
#### **#Bounding Volume Hierarchy**

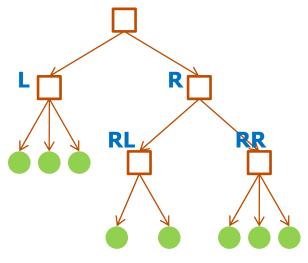


#### **\*\*Bounding Volume Hierarchy**

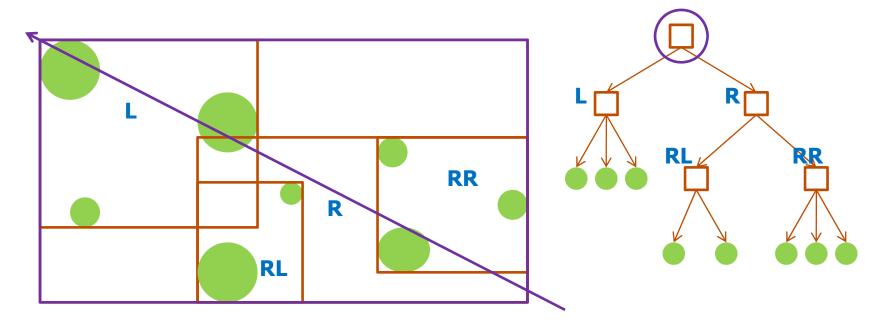


#### **\*\*Bounding Volume Hierarchy**

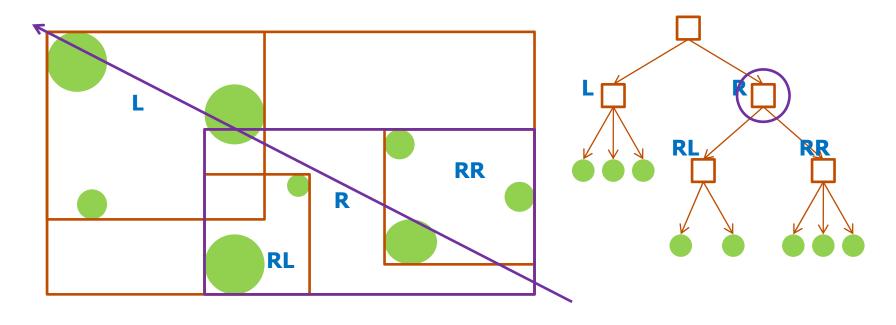




#### **ж** Траверс на CPU

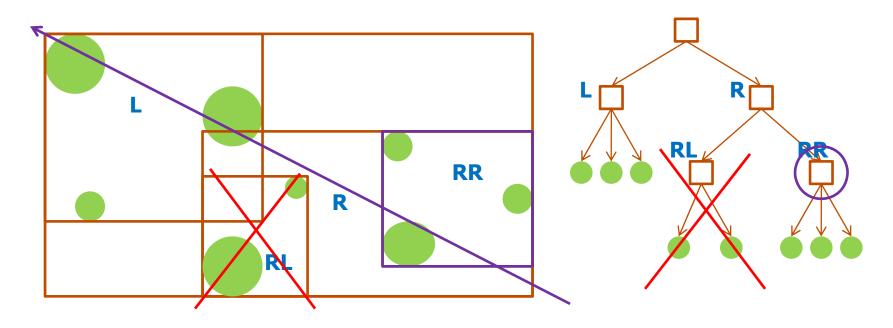


#### **ж** Траверс на CPU



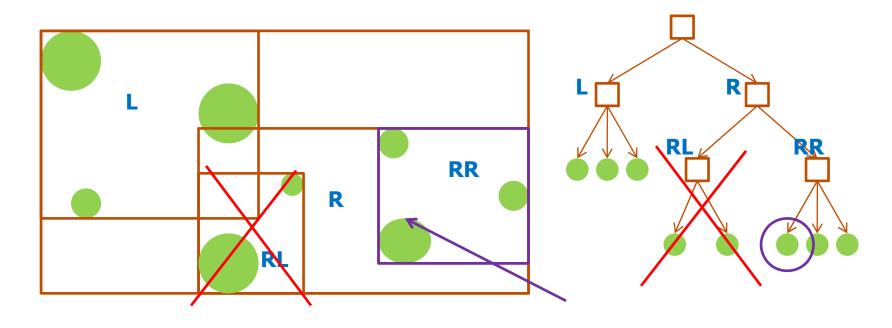
Стек: L

#### **ж** Траверс на CPU

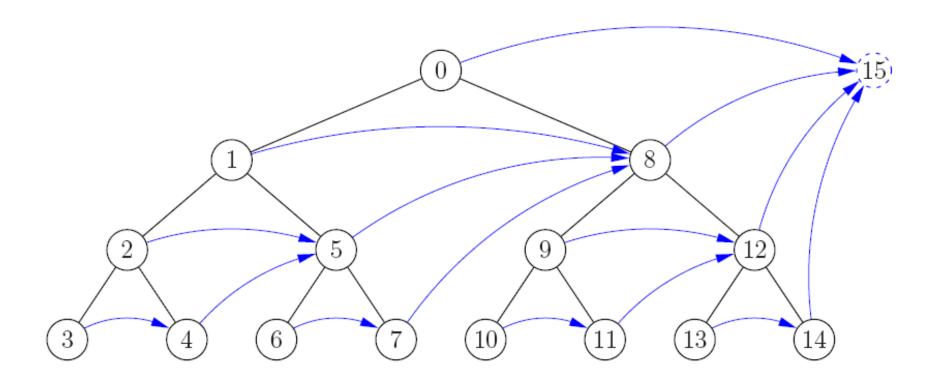


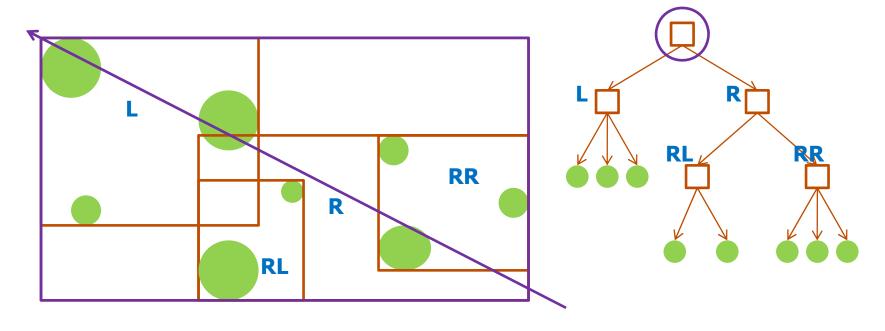
Стек: L

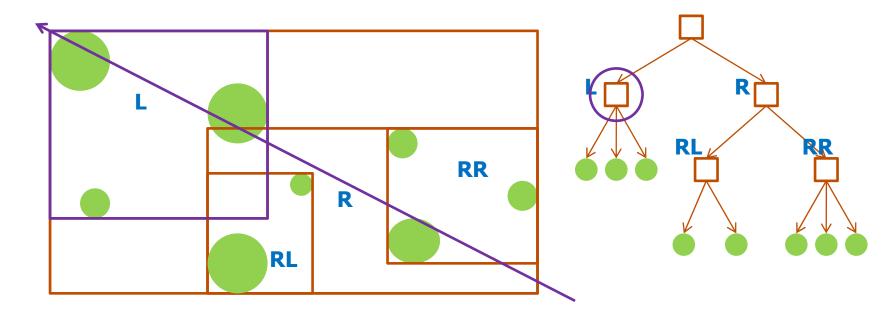
#### **ж** Траверс на CPU

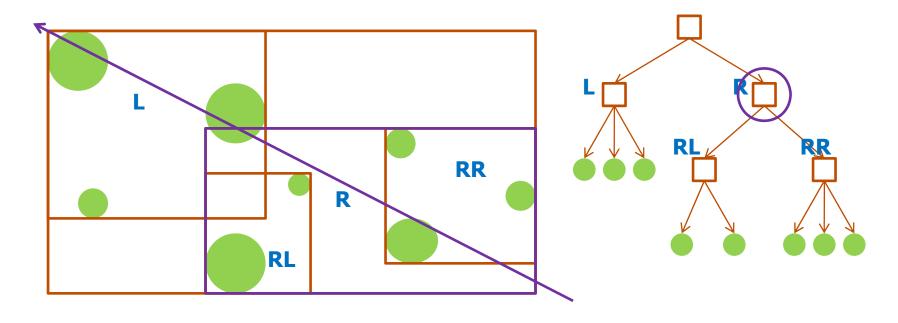


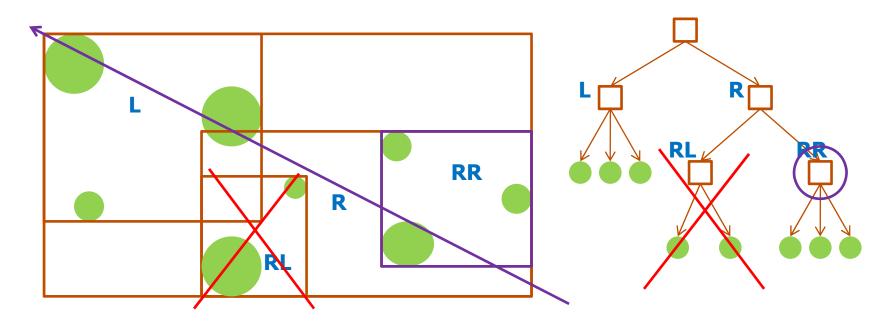
Стек: L

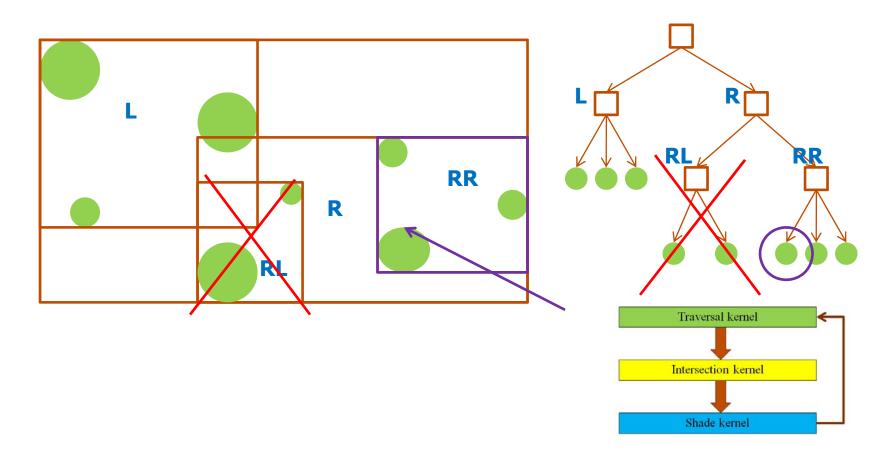






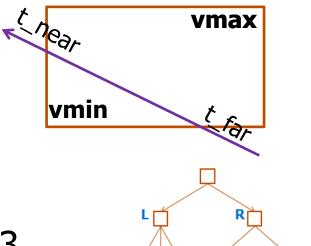




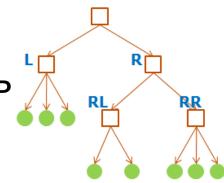


- - № Бокс 6 регистров

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

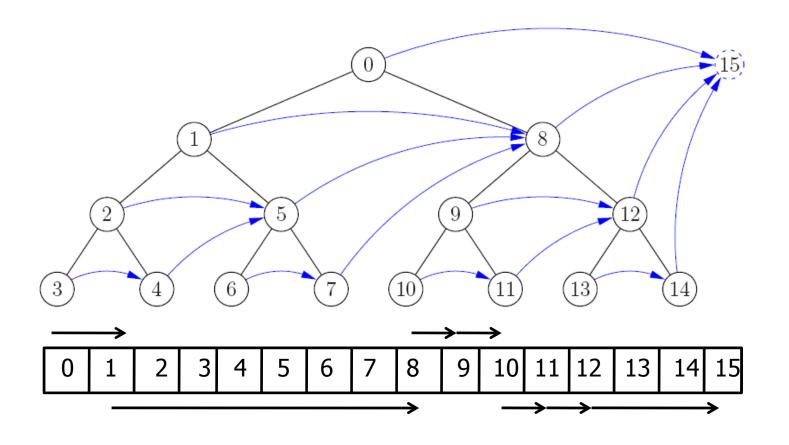


- - 24 mad-а покрывают латентность текстурной памяти

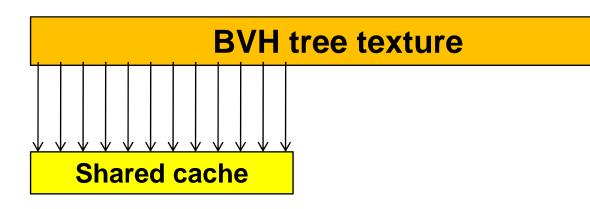


- 1. Стек на локальной памяти
  - Локальная память это не так медленно, как может показаться
- 2. Бесстековый алгоритм
  - Перебираем массив всегда строго слева направо

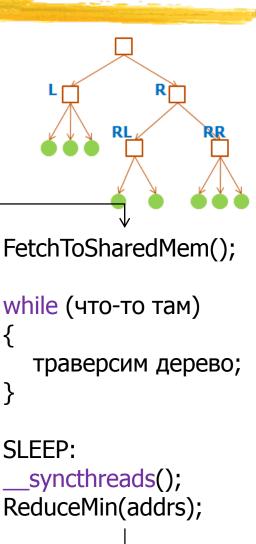
#### 



#Бесстековый траверс на CUDA

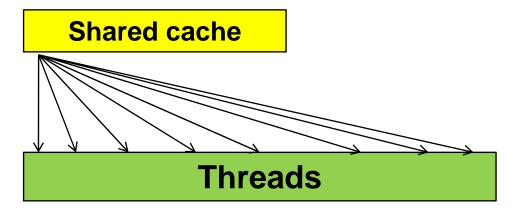


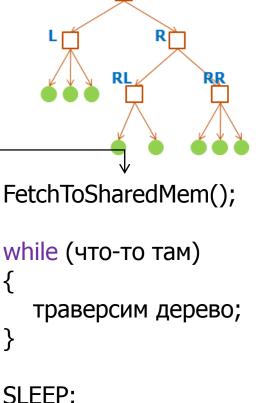
**Threads** 



#Бесстековый траверс на CUDA





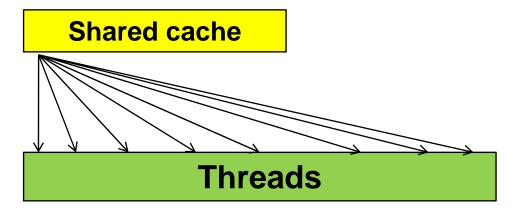


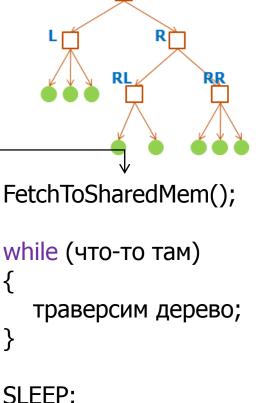
\_syncthreads();

ReduceMin(addrs);

#Бесстековый траверс на CUDA





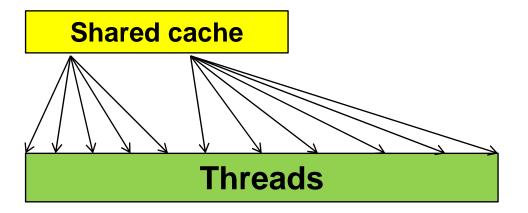


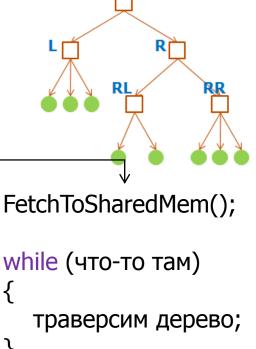
\_syncthreads();

ReduceMin(addrs);

#Бесстековый траверс на CUDA







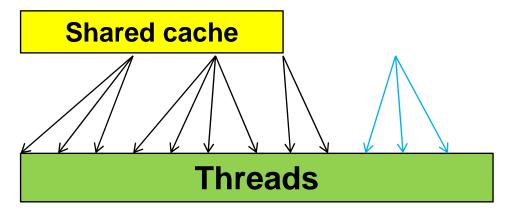
SLEEP:

\_syncthreads();

ReduceMin(addrs);

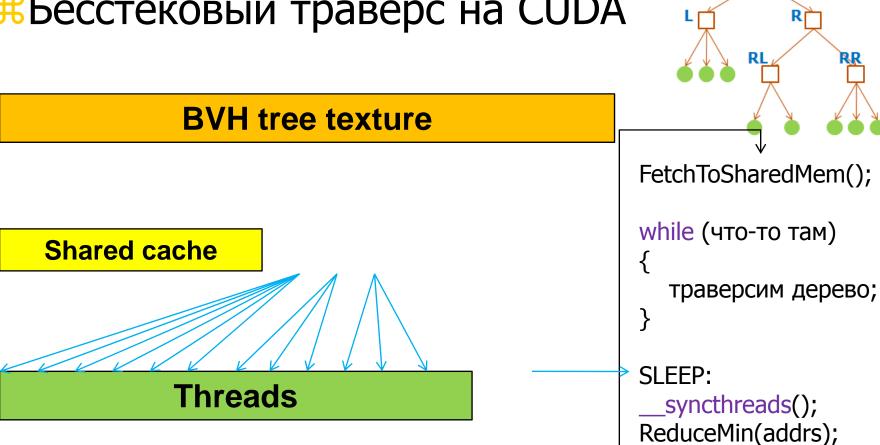
#Бесстековый траверс на CUDA



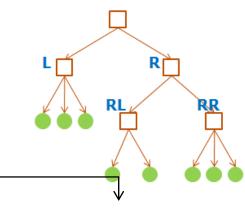


```
FetchToSharedMem();
while (что-то там)
  траверсим дерево;
SLEEP:
 _syncthreads();
ReduceMin(addrs);
```

#Бесстековый траверс на CUDA



#Бесстековый траверс на CUDA



#### **BVH** tree texture

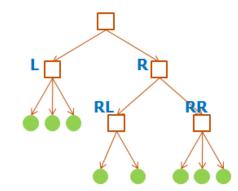
FetchToSharedMem();

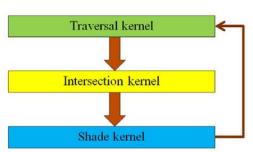
```
while (что-то там)
{
    траверсим дерево;
}
```

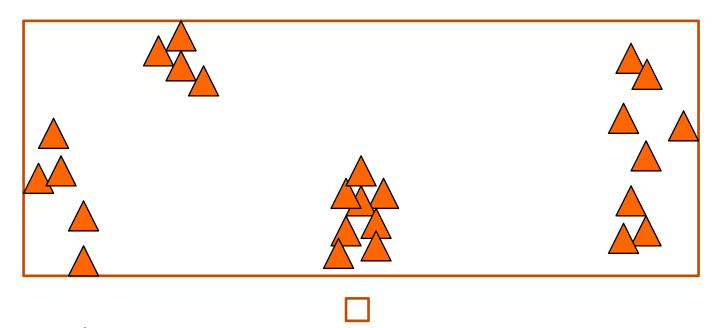
SLEEP:

```
__syncthreads();
ReduceMin(addrs);
```

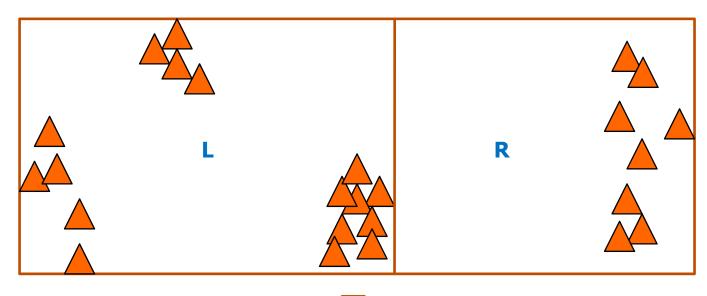
- #Бесстековый траверс на CUDA
  - Ядро уместилось в 20 регистров
  - № Узел BVH 32 байта
  - Shared кэш слишком маленький
  - На больших сценах эффективность падает из-за частых промахов
  - Плохо для некогерентных лучей
  - Лишние пересечения



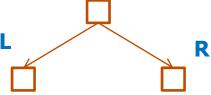


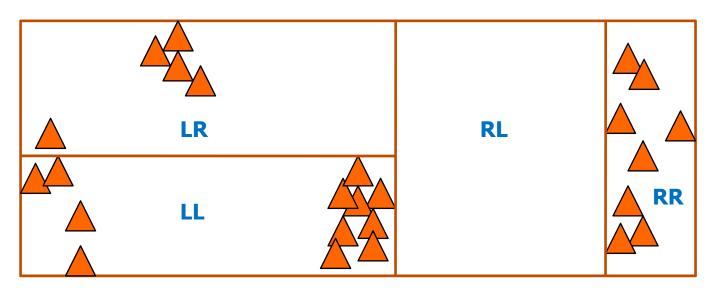


```
struct KdTreeNode
{
    float split;
    uint leftOffset: 29;
    uint splitAxis: 2;
    uint leaf: 1;
};
```

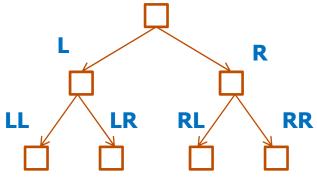


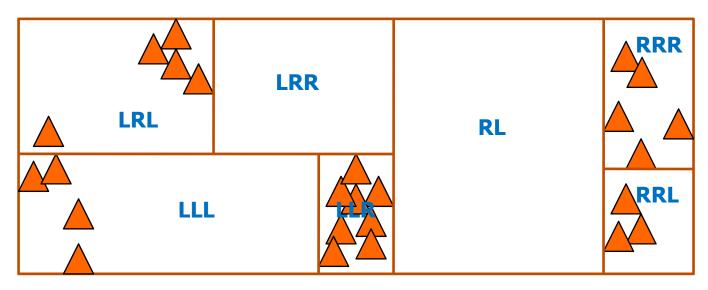
```
struct KdTreeNode
{
    float split;
    uint leftOffset: 29;
    uint splitAxis: 2;
    uint leaf: 1;
};
```



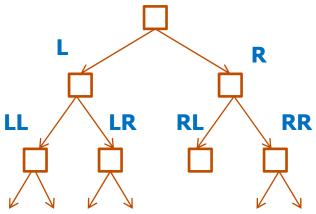


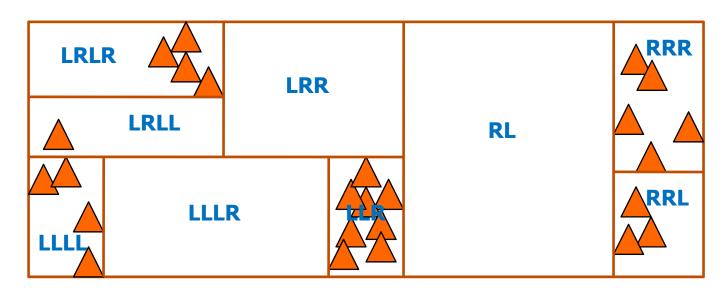
```
struct KdTreeNode
{
    float split;
    uint leftOffset: 29;
    uint splitAxis: 2;
    uint leaf: 1;
};
```



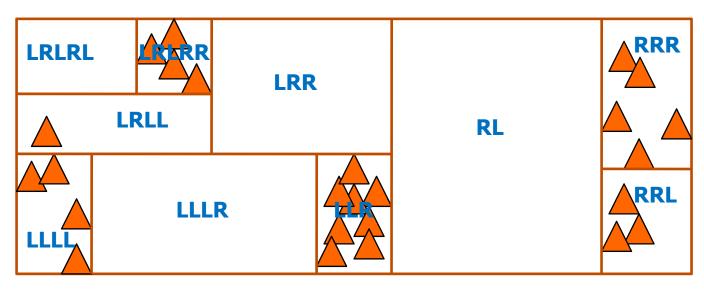


```
struct KdTreeNode
{
    float split;
    uint leftOffset: 29;
    uint splitAxis: 2;
    uint leaf: 1;
};
```

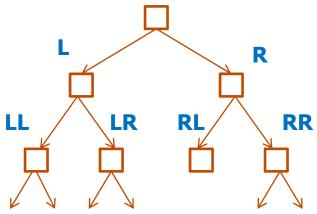




```
struct KdTreeNode
{
    float split;
    uint leftOffset: 29;
    uint splitAxis: 2;
    uint leaf: 1;
};
```



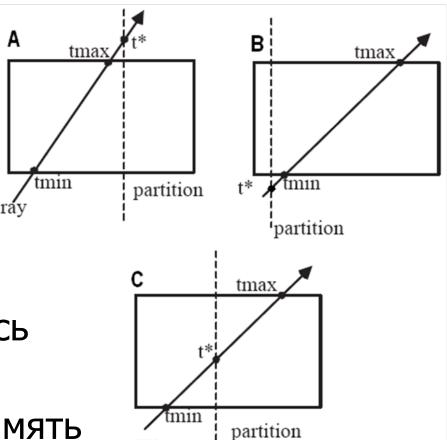
```
struct KdTreeNode
{
    float split;
    uint leftOffset: 29;
    uint splitAxis: 2;
    uint leaf: 1;
};
```



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

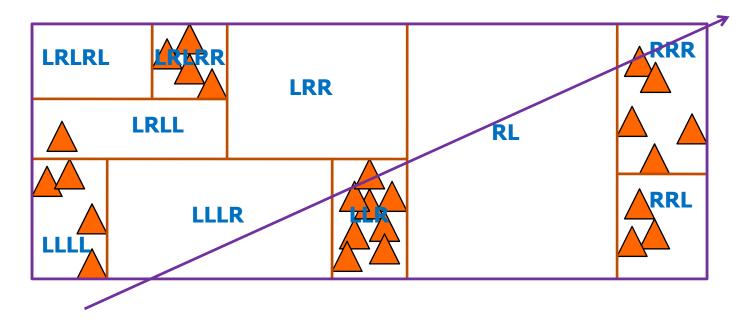
  - $\triangle$  t, tmin, tmax -3
  - node − 2

  - Стек: локальная память



ray

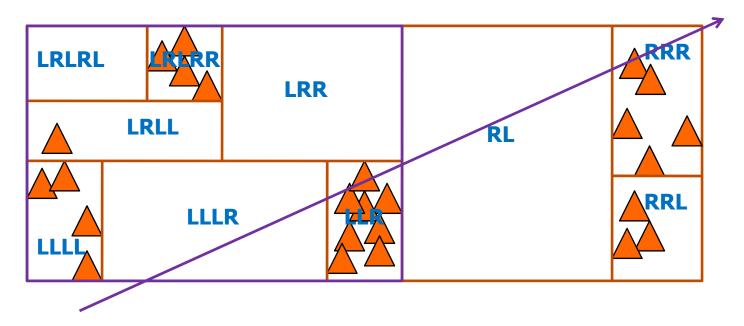
#### **Ж**Алгоритм траверса



Стек:

Текущий узел:

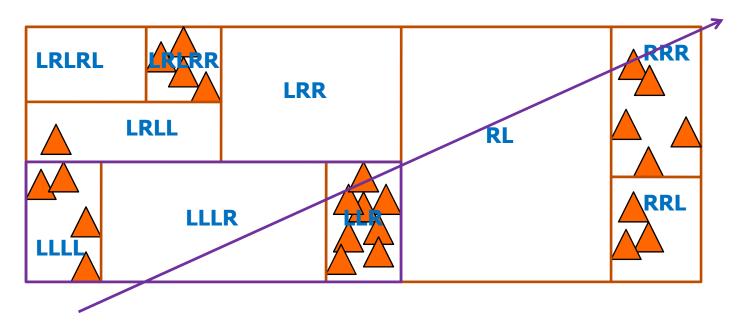
#### **Ж**Алгоритм траверса



Стек: R

Текущий узел: L

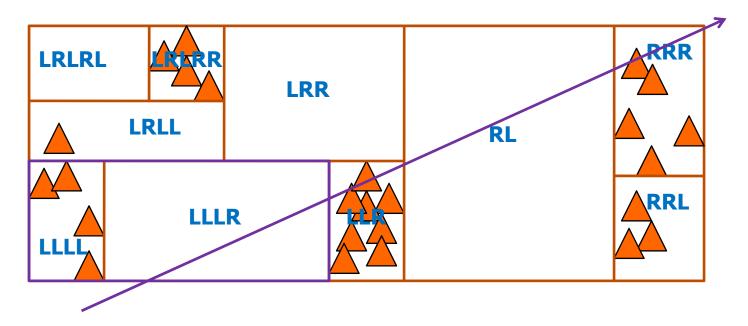
#### **Ж**Алгоритм траверса



Стек: R

Текущий узел: LL

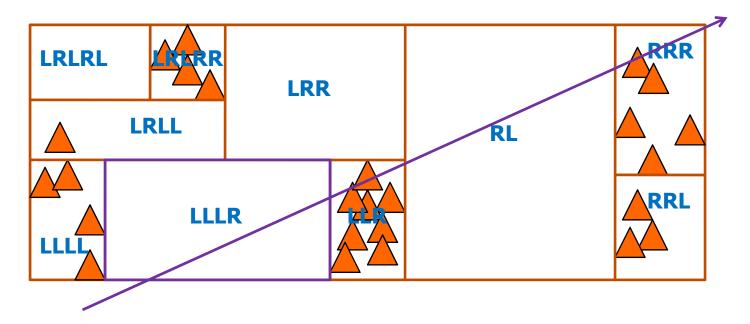
#### **Ж**Алгоритм траверса



Стек: LLR, R

Текущий узел: LLL

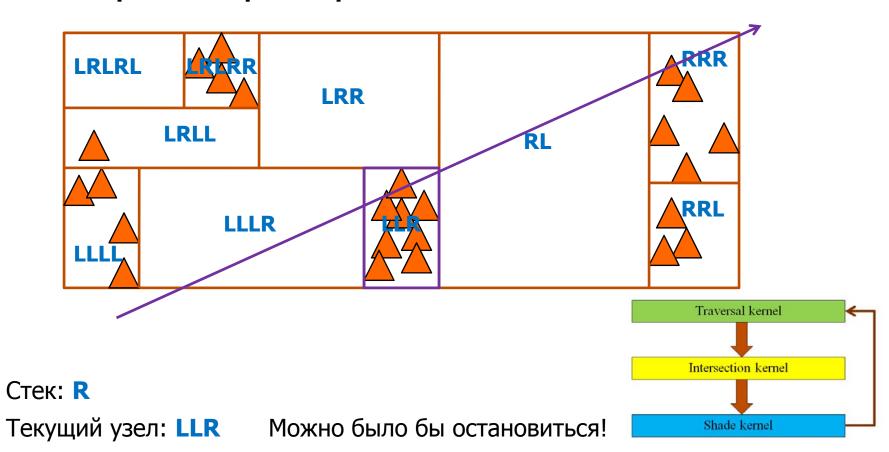
#### **Ж**Алгоритм траверса



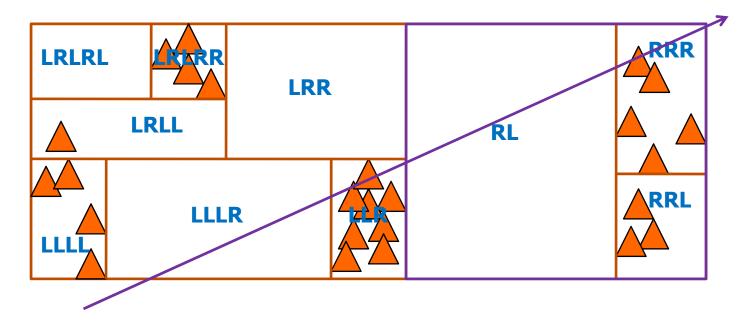
- Стек: LLR, R

Текущий узел: LLLR

#### **Ж**Алгоритм траверса



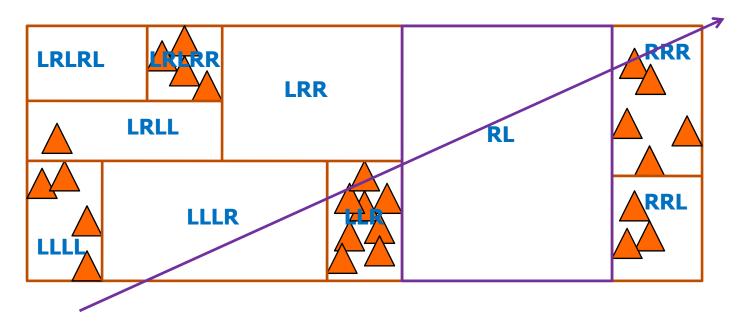
#### **Ж**Алгоритм траверса



Стек:

Текущий узел: R

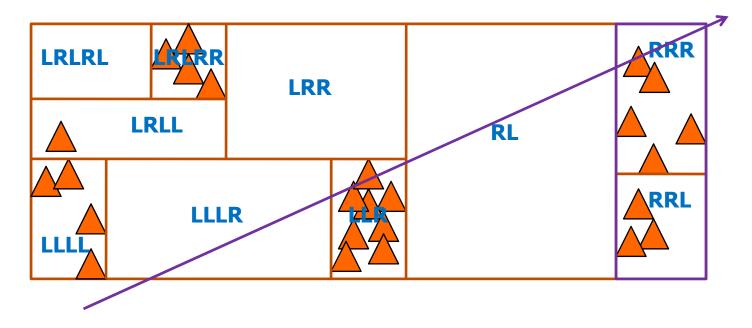
#### **Ж**Алгоритм траверса



- Стек: **RR** 

Текущий узел: RL

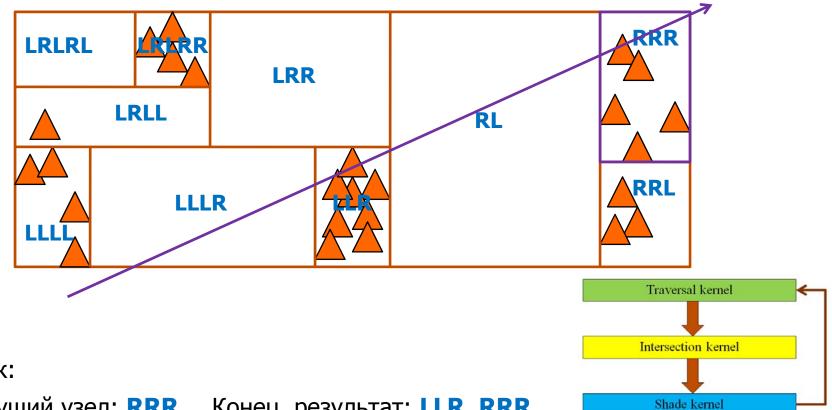
#### **Ж**Алгоритм траверса



Стек:

Текущий узел: RR

#### **Ж**Алгоритм траверса

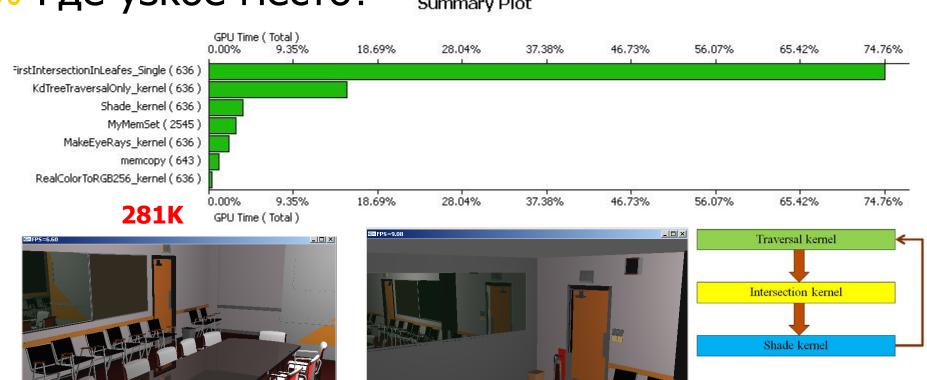


Стек:

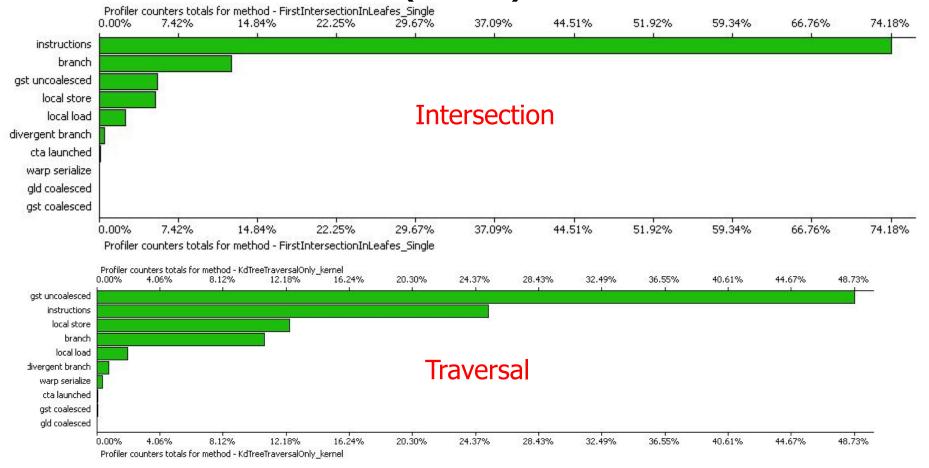
Текущий узел: RRR Конец, результат: LLR, RRR

#### **ж** Где узкое место?

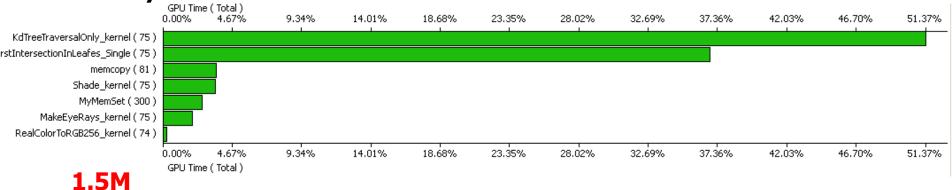
#### Summary Plot



#### # Conference Room (281K)

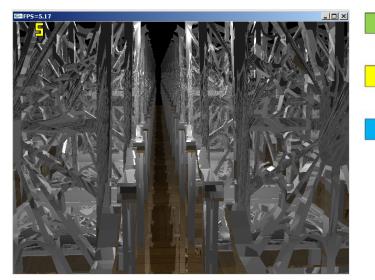


#### **ж** Где узкое место?



\_ U ×

FP5=13.16

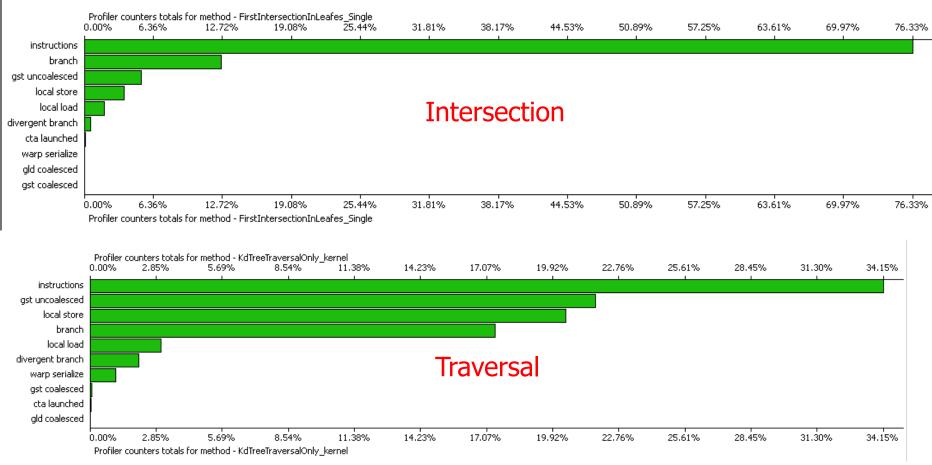


Traversal kernel

Intersection kernel

Shade kernel

### **Ж** Стулья (1.5М)



### Производительность



#### kd-tree vs BVH на CUDA

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

# Грабли

- **ж** Что делает этот код?
- **#** Почему он так написан?

```
const float3 V[2];
register int sign = (r.dir.x < 0);
register float tmin = ( V[sign].x - r.pos.x) / r.dir.x;
register float tmax = ( V[1-sign].x - r.pos.x) / r.dir.x;</pre>
```

#### **ж** А как надо?

```
if (r.dir.x < 0);
    tmin = ( V[1].x - r.pos.x) / r.dir.x;
else
    tmin = ( V[0].x - r.pos.x) / r.dir.x;</pre>
```

# Грабли

#### **ж** Очень удобно?

```
struct VECTOR
struct VECTOR
                                float M[3];
  union
                                IDH_CALL float& x() { return M[0]; }
     struct
                                IDH_CALL float& y() { return M[1]; }
                                IDH_CALL float& z() { return M[2]; }
        float x,y,z;
                                IDH_CALL float x() const { return M[0]; }
     float M[3];
                                IDH_CALL float y() const { return M[1]; }
  };
                                IDH_CALL float z() const { return M[2]; }
};
                             };
```

#define IDH\_CALL inline \_\_device\_\_ \_\_host\_\_

#### Ссылки

- ★ <a href="http://ray-tracing.ru">http://ray-tracing.ru</a>
- # http://www.nvidia.com/object/cuda home.html#state=detailsOpen ;aid=73926e40-99f0-11dd-ad8b-0800200c9a66
- # <a href="http://www.mpi-inf.mpg.de/~guenther/BVHonGPU/index.html">http://www.mpi-inf.mpg.de/~guenther/BVHonGPU/index.html</a> (BVH packet traversal + CUDA)

